

# Fast Identification of Biological Pathways Associated with a Quantitative Trait Using Group Lasso with Overlaps

BY MATT SILVER, GIOVANNI MONTANA<sup>1</sup>  
& ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

Department of Mathematics  
Imperial College London  
London, SW7 2AZ, UK

## Abstract

Where causal SNPs (single nucleotide polymorphisms) tend to accumulate within biological pathways, the incorporation of prior pathways information into a statistical model is expected to increase the power to detect true associations in a genetic association study. Most existing pathways-based methods rely on marginal SNP statistics and do not fully exploit the dependence patterns among SNPs within pathways. We use a sparse regression model, with SNPs grouped into pathways, to identify causal pathways associated with a quantitative trait. Notable features of our pathways group lasso with adaptive weights (P-GLAW) algorithm include the incorporation of all pathways in a single regression model, an adaptive pathway weighting procedure that accounts for factors biasing pathway selection, and the use of a bootstrap sampling procedure for the ranking of important pathways. P-GLAW takes account of the presence of overlapping pathways and uses a novel combination of techniques to optimise model estimation, making it fast to run, even on whole genome datasets. In a comparison study with an alternative pathways method based on univariate SNP statistics, our method demonstrates high sensitivity and specificity for the detection of important pathways, showing the greatest relative gains in performance where marginal SNP effect sizes are small.

## 1 Introduction

The mixed success of attempts to identify genetic variants that account for a large part of the heritability of common disease has focussed attention on the need to develop new methodological approaches to the analysis of GWAS data. A number of factors that might explain this 'missing heritability' have been suggested, including the failure of many current models to capture the presence of gene-gene and gene-environment interactions, of multiple SNPs with small effect and of rare variants (Manolio et al., 2009; Goldstein, 2009). One promising approach uses prior information on functional structure present within the genome to group genes and associated SNPs into gene sets or pathways. The motivation here is that genes do not work in isolation, but instead work together through their effect on molecular networks and cellular pathways. The hope is that by jointly considering the effects of multiple SNPs or genes within a biological pathway, significant associations might be identified that would otherwise be missed when considering markers individually (Wang et al., 2010). First developed in the context of gene expression studies (Mootha et al., 2003), pathways-based methods have more recently been extended to the analysis of GWAS data (Holmans et al., 2009; Luo et al., 2010; Lango Allen et al., 2010; Lambert et al., 2010). This has led to the identification of putative causal pathways for a number of diseases including Parkinson's Disease (Lesnick et al., 2007), Crohn's Disease (Wang et al., 2009b) and rheumatoid arthritis (Eleftherohorinou et al., 2011). As well as offering the potential for increased statistical power, pathways-based genetic association studies (PGAS) can aid the biological interpretation of results through the identification of causal pathways, and may also facilitate comparisons between studies genotyping different variants that nonetheless map to common pathways (Ma and Kosorok, 2010; Cantor et al., 2010).

The majority of existing PGAS methods begin with a univariate test of association, in which individual SNPs are scored according to their degree of association with disease status or a quantitative trait. Various techniques are then used to combine these univariate statistics into pathway

<sup>1</sup>Corresponding author. Email: [g.montana@ic.ac.uk](mailto:g.montana@ic.ac.uk)

scores. For example, the GenGen method (Wang et al., 2007) first ranks all genes according to the value of the highest-scoring SNP within 500kb of each gene. Pathway significance is then assessed by determining the degree to which high-ranking genes are over-represented in a given gene set, in comparison with the genomic background. The PLINK toolkit (Purcell et al., 2007) also features a ‘set-based test’, in which pathway significance is measured by taking the average, marginal p-value of a pre-determined maximum number of ‘uncorrelated’ SNPs within the pathway. Here, uncorrelated SNPs are defined as those whose pairwise linkage disequilibrium (LD) is below a certain threshold value. As a final step, where more than one pathway is considered a correction for multiple testing is generally made.

In contrast to univariate, ‘one SNP at a time’ methods, multivariate or multi-locus methods allow all SNPs to be considered in the model at the same time, which can aid the identification of weak signals while diminishing the importance of false ones. One such approach consists of fitting a penalised, multivariate regression model, in which a subset of SNPs is selected by imposing a penalty on some suitably selected norm of the regression coefficients, as in Lasso regression (Tibshirani, 1996). This approach has been shown to yield higher statistical power, compared to more common ‘mass univariate linear models’, especially with multivariate and high-dimensional quantitative traits (Vounou et al., 2010). Several other studies have demonstrated the advantages of this approach for the detection of genetic associations. For example, Wu et al. (2009) use penalized logistic regression to select SNPs in a case-control study, and analyse two-way and higher-order SNP-SNP interactions. Hoggart et al. (2008) propose a similar method for SNP selection in a Bayesian context.

A number of penalized regression techniques that allow prior information on the relationship between SNP markers to be incorporated into the model selection process have recently been proposed. For example, Zhou et al. (2010) group SNPs into genes, and utilise a useful property of the group lasso (Yuan and Lin, 2006) to aid the detection of rare variants within genes. The GRASS method (Chen et al., 2010) begins by characterising within-gene variation as ‘eigenSNPs’, obtained by principal component analysis (PCA). A combination of lasso and ridge regression, followed by permutations is then used to measure significance for a single pathway. Finally, Zhao et al. (2011) use a combination of PCA and lasso regression to identify a subset of genes within a candidate pathway, followed by permutations to measure pathway significance. Once again this method considers one pathway at a time.

The search for SNPs, or quantitative trait loci (QTL) influencing quantitative traits is gaining momentum as a potentially more powerful way to study the underlying causes of complex disease (Plomin et al., 2009). In the emerging field of neuroimaging genetics for example, in which we have a particular interest, quantitative data in the form of MRI or PET scans serve as a type of intermediate phenotype in the study of complex disorders such as Alzheimer’s Disease (AD) or schizophrenia (Bigos and Weinberger, 2010). We use genotype data from the Alzheimers Disease Neuroimaging Initiative (ADNI) dataset in this analysis.

Our focus here is on the identification of biological pathways associated with a quantitative trait. Our assumption is that where causal SNPs are enriched in a pathway, the use of a regression model that selects SNPs that are grouped into pathways will have increased power, compared to a more traditional approach in which SNPs are considered one at a time. We also seek a true, multivariate model which includes all mapped pathways at the same time. The hope is that this will confer some of the benefits, in terms of detecting weaker signals and diminishing false positives, described earlier. To achieve these ends, we use a modified version of the group lasso (GL) with SNPs grouped into pathways, and develop a fast estimation algorithm applicable to the case of non-orthogonal groups. In order to rank pathways, we use a bootstrap sampling procedure to rank pathways in decreasing order of importance. We face a number of challenges in applying GL to SNP and pathway data for the identification of implicated pathways. These include the fact that pathways overlap, since many SNPs map to multiple pathways; the problem of selection bias, that is the tendency of the model to select pathways having specific statistical properties irrespective of their association with phenotype; and the sheer scale of SNP datasets, making efficient estimation a necessity.

We have found that the issue of overlapping pathways receives surprisingly little attention in

the PGAS literature, given that the presence of overlaps might be expected to have a significant impact on the results of any PGAS analysis. For example, variation in the number and distribution of causal SNPs with respect to genes that overlap multiple pathways will affect the number of pathways defined to be ‘causal’, and different PGAS methods will be affected by such variation in different ways. Additionally, the inclusion of multiple pathways in a single GL regression model presents a particular problem, since GL in its original formulation will not select pathways in the manner that we would wish. To account for this we employ a variable expansion procedure, originally proposed in the context of microarray data analysis by Jacob et al. (2009), that ensures that overlapping SNPs enter the regression model separately, for each pathway that they map to.

A number of factors may bias PGAS results, exaggerating pathway significance and giving rise to inflated numbers of false positives. Depending on the methods used, and the underlying disease-causing mechanism, such factors are likely to include pathway size (measured in number of SNPs and/or genes), and the extent and distribution of pathway LD. Common strategies employed by existing methods to reduce this bias include the use of permutation (of genes or phenotypes), and dimensionality reduction techniques such as PCA (Fridley and Biernacka, 2011; Wang et al., 2010). We propose a procedure that reduces bias by adjusting pathway weightings in the regression model according to the empirical bias in pathway selection frequencies obtained by fitting the GL model with a null response.

One potential drawback of using a regression model in the analysis of genetic data is the typically very large number of predictors (here SNPs) that must be analysed. While the use of penalized regression techniques at least makes the problem tractable when the number of predictors vastly exceeds sample size, the very large matrix calculations required can still make model estimation computationally infeasible. To address this, we combine a number of techniques that speed up the estimation process including the use of an ‘active set’ of predictors, a Taylor approximation of the GL penalty and efficient computation of pathway block residuals. The final estimation algorithm, which we call ‘Pathways Group Lasso with Adaptive Weights’ (P-GLAW), is sufficiently fast to obviate the need either to undertake a preliminary stage of dimensionality reduction, or to consider pathways individually.

We evaluate our method’s performance in a Monte Carlo (MC) simulation study, using real genetic and pathway data with quantitative phenotypes simulated under an additive genetic model. We consider a range of scenarios with different causal SNP distributions and effect sizes. We feel the use of real genotype and pathway data is crucial, so as to capture the complex distributions of gene size and number within a pathway, together with SNP LD patterns and overlaps between pathways, all of which may have a significant effect on pathway ranking performance. To our knowledge, this is the first such PGAS power study using GL with real SNP and pathway data. The evaluation of GL pathway ranking performance however presents a number of challenges. Firstly, as described above, variation in the number of causal pathways due to overlaps must be taken into account when evaluating performance over multiple MC simulations. Secondly, we require a means of evaluating the degree to which causal pathways are represented amongst the top ranks. Thirdly, since GL performs variable selection, not all causal pathways may be ranked, and ranking performance measures must reflect this. To address these issues we devise a battery of measures that aim to capture different aspects of ranking performance. Finally, we compare our method’s performance with another common PGAS method, derived from univariate SNP statistics.

The article is organised as follows. Section 2 describes the GL model; our strategy for dealing with overlapping pathways, model estimation and speed-ups; our proposed bias-adjusted pathway weighting update procedure; our strategy for ranking pathways using a resampling procedure, and our proposed ranking performance measures. In Section 3 we describe the real biological data sets used in the experiments, the SNP to pathway mapping process, and the simulation framework used to evaluate both methods under consideration. The results from these simulation studies are provided in Section 4, and we conclude in Section 5 with a discussion and final remarks.

## 2 Methods

### 2.1 The group lasso for pathway selection

We assume  $N$  unrelated individuals genotyped at  $P$  SNPs, each with a univariate quantitative trait  $y_i$ , for  $i = 1, \dots, N$ . For an individual  $i$ , we denote by  $x_{ij}$  the minor allele count for SNP  $j$ , for  $j = 1, \dots, P$ , and arrange these counts in an  $(N \times P)$  design matrix  $\mathbf{X}$ . Quantitative phenotypes are arranged in an  $(N \times 1)$  column vector  $\mathbf{y}$ , and will be treated as quantitative responses in a regression model.

We initially consider the situation where SNPs are partitioned into  $L$  mutually exclusive pathways, or groups. Each group  $\mathcal{G}_l$ , for  $l = 1, \dots, L$ , is a subset of  $\{1, 2, \dots, P\}$  of cardinality  $S_l$ , containing the indices  $l_1, l_2, \dots, l_{S_l}$  of the SNPs that belong to it, such that  $\mathcal{G}_l \cap \mathcal{G}_{l'} = \emptyset$  for any  $l \neq l'$ . We denote by  $\mathcal{G} = \{1, \dots, P\}$ , the set of all SNP indices. We denote by  $\mathcal{S} \subset \{1, \dots, P\}$  the subset of SNPs that are *causal*, that is the SNPs influencing  $y$ , and additionally denote the cardinality of  $\mathcal{S}$  by  $S$ . Accordingly, we denote by  $\mathcal{C} \subset \{1, 2, \dots, L\}$  the subset of causal pathways containing one or more SNPs in  $\mathcal{S}$ , having cardinality  $|\mathcal{C}|$ . We denote the complement of  $\mathcal{C}$  by  $\mathcal{C}'$ . We also assume that  $|\mathcal{C}| \ll L$ , so that only a small proportion of all pathways are causal. Finally, we assume that  $y$  can be optimally predicted, in the least squares sense, by a linear combination of allele counts corresponding to SNPs in pathway  $\mathcal{G}_l$ , where  $l$  belongs to the set  $\mathcal{C}$ .

We denote the vector of SNP regression coefficients  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P) \in \mathbb{R}^P$ , and the parameter vector corresponding to SNPs in pathway  $\mathcal{G}_l$  only as  $\boldsymbol{\beta}_l = (\beta_{l_1}, \dots, \beta_{l_{S_l}}) \in \mathbb{R}^{S_l}$ . Under these assumptions, one or more elements of each  $\boldsymbol{\beta}_l$  for  $l \in \mathcal{C}$  are expected to be non-zero, whereas all the regression coefficients associated with SNPs that do not belong to  $\mathcal{C}$  will be zero, that is  $\boldsymbol{\beta}_l = \mathbf{0}$  for  $l \in \mathcal{C}'$ . For example, for a single causal pathway  $\mathcal{G}_l$  with causal SNPs  $\{a, b\}$  in  $\mathcal{S}$ , the sparsity pattern might look like

$$\boldsymbol{\beta} = \{ \underbrace{(0, \dots, 0)}_{\mathcal{G}_1}, \dots, \underbrace{(0, \dots, \beta_{l_a}, 0, \dots, \beta_{l_b}, 0, \dots, 0)}_{\mathcal{G}_l}, \dots, \underbrace{(0, \dots, 0)}_{\mathcal{G}_L} \}.$$

A suitable regression model that would enforce the assumed block sparsity pattern above is the group lasso (GL) (Yuan and Lin, 2006), in which estimates for  $\boldsymbol{\beta}$  are obtained by minimising the penalised least squares function

$$(1) \quad f(\boldsymbol{\beta}) = \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{l=1}^L w_l \|\boldsymbol{\beta}_l\|_2$$

with respect to  $\boldsymbol{\beta}$ , where  $\|\cdot\|_2$  denotes the  $\ell_2$  (Euclidean) norm and  $w_l$  is a pathway weighting factor for group  $l$ . Sparsity at the pathway level is encouraged through the imposition of an  $\ell_1$  lasso penalty on  $\|\boldsymbol{\beta}_l\|_2$ , which ensures that SNPs belonging to pathways not selected by the model have zero regression coefficients. For selected pathways, i.e. those with  $\boldsymbol{\beta}_l \neq \mathbf{0}$ , SNP coefficients tend to shrink, through the imposition of a ridge-type penalty on  $\|\boldsymbol{\beta}_l\|_2$ . The degree of sparsity is controlled by the regularisation parameter,  $\lambda$ , such that the number of pathways selected by the model increases with decreasing  $\lambda$ . For a given  $\lambda$ , the block sparsity pattern is determined both by the data ( $\mathbf{y}$  and  $\mathbf{X}$ ), and by the distribution of pathway weights,  $\mathbf{w} = (w_1, \dots, w_L)$ , such that an increase in  $w_l$  means that pathway  $l$  is less likely to be selected, whereas a decrease in  $w_l$  will have the opposite effect.

The GL optimisation problem associated with minimising the objective function (1) is convex, and can be solved using coordinate descent methods. Problems arise however in the situation where pathways overlap, that is when a SNP is allowed to belong to more than one pathway, so that  $\mathcal{G}_l \cap \mathcal{G}_{l'} \neq \emptyset$  for some  $l \neq l'$ . Firstly, where groups overlap, the penalty term in (1) is no longer separable into groups, since the same SNPs occur in multiple pathways, and convergence using coordinate descent is no longer guaranteed (Tseng and Yun, 2009). Secondly, if we wish to be able to select pathways independently, GL is unable to do this. We illustrate this last point using a simple example in Fig. 1 A, where we consider only three pathways,  $\mathcal{G}_1, \mathcal{G}_2$  and

$\mathcal{G}_3$ , two of which overlap. As a consequence of this, pathway parameter vectors  $\beta_1$  and  $\beta_2$  also overlap, since they have a number of SNPs in common (shaded dark grey). If a shared SNP is selected (i.e. it has a non-zero coefficient), then both pathways to which it belongs ( $\mathcal{G}_1$  and  $\mathcal{G}_2$ ) are also selected, since their corresponding pathway parameter vectors have non-zero  $\ell_2$  norms. The GL regression model thus does not meet our requirements, since in order to be able to rank pathways in order of importance, we wish to be able to distinguish overlapping pathways and select them independently. Conversely, where shared SNPs have zero coefficients, for example in the case that  $\mathcal{G}_1$  is not selected in the model, then these SNPs will have zero coefficients in each and every pathway to which they belong (here  $\mathcal{G}_1$  and  $\mathcal{G}_2$ ). Hence SNPs retained in the model are necessarily drawn from the complement of a union of (unselected) pathways. We instead require retained SNPs to be drawn from a union of (selected) pathways, so that a SNP driving selection in one pathway may still have a zero coefficient in another.

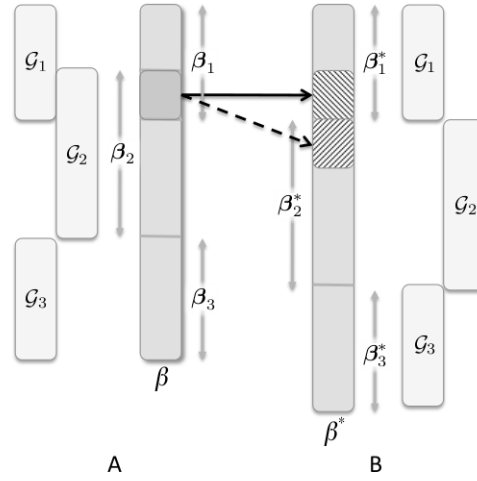


Figure 1: The problem of overlapping pathways: here there are three pathways,  $\mathcal{G}_1, \mathcal{G}_2$  and  $\mathcal{G}_3$ , two of which overlap. A: Standard formulation. Pathway parameter vectors  $\beta_1$  and  $\beta_2$  overlap, since they have SNPs in common (shaded dark grey). Where an overlapping SNP has a non-zero coefficient, only  $\mathcal{G}_3$ , can be selected independently. B: Formulation with duplicated SNPs. An expanded parameter vector,  $\beta^*$ , is created by duplicating overlapping SNPs (dotted line).  $\beta_1^*$  and  $\beta_2^*$  now enter the model separately, so that pathways can be selected independently.

Jacob et al. (2009) propose one possible solution to the problem of overlapping predictors in a similar context, motivated by the analysis of gene expression data. The essence of this method is to create duplicate, dummy SNPs, so that SNPs belonging to more than one pathway enter the model separately (see Fig. 1 B). The process works as follows. An expanded design matrix is formed from the column-wise concatenation of the  $L$  sub-matrices of size  $(N \times S_l)$ , that is  $\mathbf{X}_{\mathcal{G}_l} = \{x_{ij}\}$  with  $i = 1, \dots, N$  and  $j \in \mathcal{G}_l$ , to form the expanded design matrix  $\mathbf{X}^* = [\mathbf{X}_{\mathcal{G}_1}, \mathbf{X}_{\mathcal{G}_2}, \dots, \mathbf{X}_{\mathcal{G}_L}]$  of size  $(N \times P^*)$ , where  $P^* = \sum_l S_l$ . The corresponding parameter vector,  $\beta^*$ , size  $(P^* \times 1)$ , is formed by joining the  $L$ ,  $(S_l \times 1)$  pathway parameter vectors,  $\beta_l^*$ , so that  $\beta^* = [\beta_1^{*T}, \beta_2^{*T}, \dots, \beta_L^{*T}]^T$ . The model is then able to perform pathway selection in the way that we require, and the optimisation (1), with  $\beta$  replaced by  $\beta^*$ , and  $\mathbf{X}$  replaced by  $\mathbf{X}^*$  becomes block separable, so that it can be solved by block coordinate descent. In the following sections we assume both  $\beta$  and  $\mathbf{X}$  have been expanded, but omit the  $*$  superscript for clarity. Finally, we note that where one or more SNPs in  $\mathcal{S}$  overlap multiple pathways, the corresponding number,  $|\mathcal{C}|$ , of causal pathways will increase.

## 2.2 Parameter estimation

We seek a solution,  $\hat{\beta}$ , that minimises the GL objective function (1). Where groups or pathways are disjoint, so that the penalties are separable into groups, a global solution can be obtained using block coordinate descent (BCD). Coordinate descent algorithms offer a highly efficient means of solving convex optimisation problems, and work by breaking down the optimisation into a series of univariate problems, solving the optimisation for each variable (here SNP) in turn, while holding all the others fixed, until a suitable minimum based on some stopping criterion is reached (Friedman et al., 2007). Where variables are grouped, as in GL, estimates are obtained for each pathway parameter vector,  $\beta_l$  in turn, while holding constant the current estimates for all other pathway parameter vectors,  $\hat{\beta}_m, (m \neq l)$ , and then cycling through each pathway until convergence.

Yuan and Lin (2006) derive a method for solving GL under the assumption that the group design matrices,  $\mathbf{X}_{\mathcal{G}_l}$  are orthogonal, that is  $\mathbf{X}_{\mathcal{G}_l}^T \mathbf{X}_{\mathcal{G}_l} = \mathbf{I}$ . This assumption does not hold in our case, so in the next section we derive a solution for GL in the case of non-orthogonal groups. We additionally find that GL estimation using BCD can be slow, particularly for the large datasets common to PGAS, and so in the following sections propose a number of strategies for speeding up parameter estimation.

### 2.2.1 Block coordinate descent for non-orthogonal groups

We assume that (1) is block-separable, that is the groups indexed by  $1, \dots, L$  are disjoint by construction. In our context, this is achieved by implementing the SNP duplication strategy of section 2.1. We begin by considering a single pathway  $l$ . We collect the  $N$  individual observed SNPs for a given SNP  $j$  in a column vector  $X_j = (x_{1j}, x_{2j}, \dots, x_{Nj})$ . Using this notation, we define the matrix  $\mathbf{X}_{\mathcal{G}_l} = (X_{l_1}, X_{l_2}, \dots, X_{S_l})$  containing all  $S_l$  SNPs belonging to pathway  $\mathcal{G}_l$ , and the corresponding vector of regression coefficients  $\beta_l = (\beta_{l_1}, \beta_{l_2}, \dots, \beta_{S_l})$ . We can then rewrite the objective function (1) for a single block  $l$  as a function of  $\beta_l$ ,

$$(2) \quad f(\beta_l) = \frac{1}{2} \|\hat{\mathbf{r}}_l - \sum_{j \in \mathcal{G}_l} X_j \beta_j\|_2^2 + \lambda w_l \|\beta_l\|_2$$

where  $\hat{\mathbf{r}}_l = \mathbf{y} - \sum_{m \neq l} \mathbf{X}_m \hat{\beta}_m$ . The vector  $\hat{\mathbf{r}}_l$  is the ‘partial residual’ vector for pathway  $l$ , based on the current estimates,  $\hat{\beta}_m, m \neq l$ , of the other pathway parameter vectors.

Estimates for each  $\beta_j$  are then obtained by taking partial derivatives with respect to  $\beta_j$ , that is by setting

$$(3) \quad \frac{\partial f(\beta_l)}{\partial \beta_j} = 0 \quad \text{for } j = l_1, \dots, S_l$$

Ignoring the penalty term, the partial derivative with respect to  $\beta_j$  is

$$\frac{\partial}{\partial \beta_j} \frac{1}{2} \|\hat{\mathbf{r}}_l - \sum_j X_j \beta_j\|_2^2 = -X_j^T (\hat{\mathbf{r}}_l - \sum_j X_j \beta_j)$$

We denote the partial derivative of the penalty term, by

$$s_j = \frac{\partial}{\partial \beta_j} \|\beta_l\|_2$$

so that (3) can be written as

$$(4) \quad -X_j^T (\hat{\mathbf{r}}_l - \sum_j X_j \beta_j) + \lambda w_l s_j = 0 \quad j = l_1, \dots, S_l$$

We first consider the case where  $\beta_l = \mathbf{0}$ , that is  $\beta_j = 0$ , for  $j = l_1, \dots, S_l$ . In this case  $\|\beta_l\|_2$  is not differentiable. We instead form the  $S_l$  sub-differentials,  $s_j \in [-1, 1]$ , so that

$$(5) \quad \sum_j s_j^2 \leq 1$$

The system of equations (4) can now be written

$$s_j = \frac{1}{\lambda w_l} X_j^T \hat{\mathbf{r}}_l \quad j = l_1, \dots, S_l$$

and using (5), we have

$$(6) \quad \sum_j s_j^2 = \frac{1}{\lambda^2 w_l^2} \sum_j (X_j^T \hat{\mathbf{r}}_l)^2 \leq 1.$$

Note that for (6) to be unbiased with respect to group size, a weight,  $w_l = \sqrt{S_l}$ , as proposed by Yuan and Lin (2006), can be applied. Alternatively, since

$$\sum_j (X_j^T \hat{\mathbf{r}}_l)^2 = \|\mathbf{X}_l^T \hat{\mathbf{r}}_l\|_2^2$$

we may rewrite (6) as

$$\left(\sum_j s_j^2\right)^{\frac{1}{2}} = \frac{1}{\lambda w_l} \|\mathbf{X}_l^T \hat{\mathbf{r}}_l\|_2 \leq 1,$$

so that if  $\beta_l = \mathbf{0}$

$$(7) \quad \|\mathbf{X}_l^T \hat{\mathbf{r}}_l\|_2 \leq \lambda w_l.$$

When  $\beta_l \neq \mathbf{0}$ , the minimisation of (2) can be obtained numerically, using coordinate descent, as a series of one-dimensional estimations over  $\beta_j, j = l_1, \dots, l_{S_l}$ . Friedman et al. (2010) suggest a golden section search over  $\beta_j$ , combined with parabolic interpolation. However, the number of such estimations depends on  $L$  and  $P^*$ , both of which increase with  $P$ , the latter markedly so. This can make the GL optimisation prohibitively slow, particularly for the large  $P$  typically found in PGAS. For this reason, we next describe three strategies for speeding up the estimation.

### 2.2.2 Taylor approximation of penalty

One means of speeding up the estimation for  $\beta_j$  is to use a linear or quadratic approximation of the GL  $\ell_2$  penalty (Zou and Li, 2008; Fan and Li, 2001), enabling the replacement of the multi-step numerical optimisation over  $\beta_j$  with a one-step calculation. Breheny and Huang (2009) propose the use of a Taylor approximation for a range of different estimation problems with grouped variables and we adopt this approach for our GL estimation problem. We begin by rewriting the group  $\mathcal{G}_l$  objective function (2), for a single predictor as

$$f(\beta_l | \hat{\beta}_k, k \in \mathcal{G}_l, k \neq j) = \frac{1}{2} \|\hat{\mathbf{r}}_l - \sum_k X_k \hat{\beta}_k - X_j \beta_j\|_2^2 + \lambda w_l \Gamma(\beta_l | \hat{\beta}_k)$$

where  $\Gamma(\beta_l | \hat{\beta}_k) = (c + \beta_j^2)^{\frac{1}{2}}$ , with  $c = \sum_{k \neq j} \hat{\beta}_k^2$ , and the  $\hat{\beta}_k$  are the current SNP coefficient estimates. For convenience, we rewrite this as

$$(8) \quad f(\beta_l | \hat{\beta}_k, k \neq j) = \frac{1}{2} \|\hat{\mathbf{r}} + X_j \hat{\beta}_j - X_j \beta_j\|_2^2 + \lambda w_l \Gamma(\beta_l | \hat{\beta}_k)$$

where  $\hat{\mathbf{r}} = \mathbf{y} - \sum_l \mathbf{X}_l \hat{\beta}_l$  is the total residual, using the current estimates of all SNP coefficients. We now consider the first order Taylor expansion of  $\Gamma(\beta_l | \hat{\beta}_k)$  as a function of  $x = \beta_j^2$ , about the point  $a = \hat{\beta}_j^2$

$$\Gamma(x) \simeq \Gamma(a) + \Gamma'(a)(x - a)$$

Now

$$\Gamma(x) = (c + x)^{\frac{1}{2}}$$

$$\text{and } \Gamma'(a) = \frac{1}{2(c + a)^{\frac{1}{2}}}$$

so that

$$\Gamma(x) \simeq (c + a)^{\frac{1}{2}} + \frac{x - a}{2(c + a)^{\frac{1}{2}}}$$

Substituting  $a = \hat{\beta}_j^2$ , and noting that  $(c + a)^{\frac{1}{2}} = \|\hat{\beta}_l\|_2$ , where  $\hat{\beta}_l$  denotes the current estimate of  $\beta_l$ , this gives

$$\Gamma(\beta_j^2) \simeq \hat{\beta}_l + \frac{\beta_j^2 - \hat{\beta}_j^2}{2\|\hat{\beta}_l\|_2}$$

Substituting this expression in (8), we have

$$f(\beta_l | \hat{\beta}_k, k \neq j) = \frac{1}{2} \|\hat{\mathbf{r}} + X_j \hat{\beta}_j - X_j \beta_j\|_2^2 + \lambda w_l \left[ \hat{\beta}_l + \frac{\beta_j^2 - \hat{\beta}_j^2}{2\|\hat{\beta}_l\|_2} \right]$$

Differentiating with respect to  $\beta_j$  gives

$$\begin{aligned} \left. \frac{\partial f(\beta_l)}{\partial \beta_j} \right|_{\hat{\beta}_k, k \neq j} &= -X_j^T (\hat{\mathbf{r}} + X_j \hat{\beta}_j - X_j \beta_j) + \lambda w_l \frac{\beta_j}{\|\hat{\beta}_l\|_2} \\ &= -X_j^T \hat{\mathbf{r}} - \hat{\beta}_j + \beta_j + \lambda w_l \frac{\beta_j}{\|\hat{\beta}_l\|_2} \end{aligned}$$

since  $\sum_i x_{ij}^2 = X_j^T X_j = 1$ . Rearranging terms and setting the partial derivative equal to zero, we see that the minimum is achieved when

$$(9) \quad \beta_j = \frac{X_j^T \hat{\mathbf{r}} + \hat{\beta}_j}{1 + \lambda'} \quad \text{where } \lambda' = \frac{\lambda w_l}{\|\hat{\beta}_l\|_2}$$

Where the current estimate  $\|\hat{\beta}_l\|_2 = \mathbf{0}$ , that is when group  $l$  first enters the estimation, we set  $\|\hat{\beta}_l\|_2$  to be a small positive quantity,  $\eta$ , enabling  $\beta_j$  in (9) to be estimated.

BCD proceeds by obtaining estimates for each  $\beta_j, j = l_1, \dots, S_l, 1, \dots, S_l, \dots$  until convergence within the block, and for each pathway,  $l = 1, \dots, L, 1, \dots, L, \dots$  in turn, until a stopping criterion indicating a global minimum of (1) has been satisfied. The estimation process is summarised in Box 1.

### 2.2.3 Use of pathway ‘active set’

For large  $P^*$  and  $L$ , the need for the repeated calculation of (7) to establish whether or not a particular group can enter the estimation presents a major computational bottleneck. This problem motivates another strategy providing substantial gains in computational efficiency for a range of sparse regression problems. This ‘active set’ strategy relies on the pre-selection of a subset of ‘potentially active’ predictors, or groups of predictors that are likely to be selected by the model at a given  $\lambda$  (Tibshirani et al., 2010; Roth and Fischer, 2008). The optimisation can then be run over this reduced set of variables, with a subsequent check to ensure that no other predictors should have been included in the first place. The active set procedure offers potentially dramatic speed up in execution times, particularly for very large datasets such as those found in PGAS, due to the reduced number of computations that need to be performed. In addition there are substantial savings in the amount of memory required to store data during processing, which



---

**Box 1** GL estimation algorithm using BCD

---

1. set  $\hat{\beta} = \mathbf{0}$ .
  2. For pathway  $\mathcal{G}_l, l = l_1, 2, \dots, L$ :
    - set  $\hat{\mathbf{r}}_l = \mathbf{y} - \sum_{m \neq l} \mathbf{X}_m \hat{\beta}_m$
    - If  $\|\mathbf{X}_l^T \hat{\mathbf{r}}_l\|_2 \leq \lambda w_l$ 
      - set  $\hat{\beta}_l = \mathbf{0}$
    - else
      - do
        - for  $j = l_1, \dots, S_l$ 
          - estimate  $\beta_j$  using (9)
      - end
      - until convergence of  $f(\beta_l)$  (2)
      - set  $\hat{\beta}_l = \beta_l$
    - end
  3. Repeat step 2 until (global) convergence of  $f(\beta)$ (1)
- 

can also lead to dramatic reductions in computation times with large datasets where memory is constrained.

For the GL, we begin by considering the inequality (7). For groups to enter the model we require

$$(10) \quad \|\mathbf{X}_l^T \hat{\mathbf{r}}_l\|_2 > \lambda w_l \quad l = 1, \dots, L$$

and therefore, at the first iteration, with  $\beta$  initialised to zero, a group  $\mathcal{G}_l$  enters the model if

$$(11) \quad \|\mathbf{X}_l^T \mathbf{y}\|_2 > \lambda w_l \quad l = 1, \dots, L.$$

We define the ‘active set’  $\mathcal{A}$  of potentially active groups that satisfy (11) as

$$\mathcal{A} = \{m \in \mathcal{G} : \|\mathbf{X}_m^T \mathbf{y}\|_2 > \lambda w_m\}$$

and additionally define

$$(12) \quad \lambda_{max} = \min_{\lambda} : \|\mathbf{X}_l^T \mathbf{y}\|_2 \leq \lambda w_l \quad l = 1, \dots, L$$

namely the smallest  $\lambda$  value for which the active set is empty. Note that provided  $\lambda$  is close to  $\lambda_{max}$ , then  $|\mathcal{A}| \ll L$ . Once one or more groups enter the model, not all  $\hat{\beta}_l$  will be zero and the inequality (10) will then determine which groups may enter or leave the model.

The active set procedure rests on the observation that in practice, the final set of groups selected by the model rarely includes any groups not in  $\mathcal{A}$  (Tibshirani et al., 2010). We can therefore perform the full estimation on  $\mathcal{A}$ , followed by a check of the inequality (10), to see if any additional groups not in  $\mathcal{A}$  can enter the model. If there are no additional groups, then we have the full solution. If not, then we run the full estimation again, with the additional groups satisfying (10) added to  $\mathcal{A}$ . A summary of the active set algorithm is given in Box 2.

#### 2.2.4 Efficient computation of block residuals

A further, large computational burden results from the repeated calculation of the residuals  $\mathbf{r}_l$  and  $\mathbf{r}$  in (7), (9) and (10). The computational overhead for these calculations is substantial, both because of the size of the expanded design matrix ( $N = 743$  and  $P^* = 66,085$  in the

---

**Box 2** Active set algorithm for a single  $\lambda$  value

---

1. Form the active set,  $\mathcal{A} = \{m \in \mathcal{G} : \|\mathbf{X}_m^T \mathbf{y}\|_2 > \lambda w_m\}$
2. Set  $\hat{\boldsymbol{\beta}} = \mathbf{0}$ , and solve the GL estimation at  $\lambda$ , using only the groups in  $\mathcal{A}$ :

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{y} - \sum_{m \in \mathcal{A}} \mathbf{X}_m \boldsymbol{\beta}_m\|_2^2 + \lambda \sum_{m \in \mathcal{A}} w_m \|\boldsymbol{\beta}_m\|_2$$

3. Compute the revised active set on the full dataset:

$$\mathcal{A}^+ = \{z \in \mathcal{G} : \|\mathbf{X}_z^T \hat{\mathbf{r}}_z\|_2 > \lambda w_z\}$$

```

if  $\mathcal{A}^+ / \mathcal{A} = \emptyset$ 
     $\hat{\boldsymbol{\beta}}$  is the full solution
    STOP
else
    set  $\mathcal{A} = \mathcal{A}^+$ 
    repeat 2. and 3. with the new, (expanded) active set
end

```

---

simulation study described in section 3, but substantially larger for a full PGWAS), and because of the iterative nature of the BCD algorithm, meaning that a very large number of calculations are performed. We therefore achieve one further substantial gain in computational efficiency by noting that since the blocks are separable, during BCD only the single block residual,  $\mathbf{h}_l = \mathbf{y} - \mathbf{X}_l \boldsymbol{\beta}_l$ , changes between iterations  $j = 1, \dots, S_l, 1, \dots, S_l, \dots$  within block  $l$ , and between iterations  $l = 1, \dots, L, 1, \dots, L, \dots$  across blocks. We therefore only need update  $\mathbf{h}_l$  at each iteration, with  $\mathbf{r}$  and  $\mathbf{r}_l$  updated using computationally inexpensive matrix subtractions and additions. Python code for mapping SNPs to pathways, and for analysing SNP data using PGLAW is available on request.

### 2.3 Selection bias and pathway weighting

PGAS methods derived from univariate SNP statistics are subject to various biasing factors that can influence pathway ranking under the null, where no SNPs influence the phenotypic trait,  $y$ . These factors vary from method to method, but may include the number and size of genes in a pathway, as well as LD between SNPs and genes. Such biasing factors are generally corrected through the use of permutation procedures. For example, the ‘GenGen’ method (Wang et al., 2009b), measures the degree to which pathways are enriched with high ranking genes, and is subject to bias due to variation in the number of SNPs mapped to a gene, and to differences in LD between SNPs mapped to different genes. The bias correction procedure begins by forming multiple datasets through permutation of phenotype labels. For each permuted dataset, gene scores are generated from univariate SNP statistics, and a pathway enrichment score is calculated. A normalised (bias-corrected) pathway enrichment score is then derived by comparing the distribution of pathway scores under the null with the score obtained from the unpermuted data.

Regression-based methods are similarly prone to bias, and once again the use of permutation has been proposed to correct for this, along with dimensionality reduction to extract non-redundant information. For example, with the GRASS method for case-control data (Chen et al., 2010), genetic information within each gene is first summarised as ‘eigenSNPs’, obtained through PCA. The biasing effect of gene size is once again accounted for through the generation of a null distribution, formed by permuting phenotype labels.

With the GL under the null, pathway selection will be influenced by pathway size (i.e. the number of SNPs within a pathway), since the accumulation of spurious associations in larger pathways will give rise to larger  $\|\boldsymbol{\beta}_l\|_2$  in (1). In addition, variation in dependencies between

SNPs within pathways, and to a lesser extent between pathways will give rise to corresponding variations in  $\|\beta_l\|_2$  where spurious associations arise in regions of high LD.

One way to correct for biases arising from variations in the statistical properties of different pathways or groups is through the selection of appropriate group weights  $\mathbf{w} = (w_1, \dots, w_L)$  for the objective function (1). For example, as noted before, Yuan and Lin (2006) suggest one possible choice for the pathway weighting would be

$$(13) \quad w_l = \sqrt{S_l}$$

which ensures that groups of different size are penalised equally, and so have an equal chance of being selected by the model, other things being equal (see (6)). In principle, we could follow this strategy and perhaps attempt to account for other, additional factors that may also bias pathway selection. However, there are a number of problems with this approach. Consider for example the biasing effect of dependencies between SNPs within a pathway. Where causal SNPs tag, or reside within large blocks with strong LD, the pathway ‘signal’ will be high, increasing the chance that such pathways will be selected by the model, compared with other pathways where LD is low. This biasing effect will further depend on the distribution of LD within the pathway, which will in turn depend on other factors such as the number and size of pathway genes. The precise form of any additional term(s) that should be added to (13) to account for this bias is thus unclear. Even if we were able to identify a list of potential biasing factors, and formulate bias-correcting weight adjustments for each, we are still faced with the problem that there may be other, unknown factors that contribute to the bias. We therefore choose to adopt a ‘hypothesis-free’ approach to adjusting pathway weights, which makes no assumptions about those factors which might influence pathway selection.

Consider pathway selection under the GL model (1), with  $\lambda$  tuned to select  $M$  pathways. We begin with the case  $M = 1$ . When there is no selection bias, and assuming no genetic association, a pathway  $\mathcal{G}_l$  should be randomly selected by the model according to a uniform distribution, namely with probability  $\Pi_l = 1/L$ , for  $l = 1, \dots, L$ . However, when biasing factors are present this is generally not the case, and the empirical probability distribution describing pathway selection,  $\Pi^*(\mathbf{w})$  will not be uniform. Here the dependence upon the weight vector  $\mathbf{w}$  has been made explicit, since with  $\lambda$  tuned to select a single pathway,  $\mathbf{w}$  alone determines the frequency distribution. A measure of distance between these two distributions can be obtained by computing their Kullback-Leibler (KL) divergence

$$(14) \quad D = \sum_l \Pi_l^*(\mathbf{w}) \log \frac{\Pi_l^*(\mathbf{w})}{\Pi_l}$$

where  $\Pi_l^*(\mathbf{w})$  is the empirical probability for the selection of pathway  $\mathcal{G}_l$  under the assumption of no genetic associations. When GL pathway selection is unbiased, we expect this distance to be approximately zero. Our strategy consists in adaptively adjusting all weights  $\mathbf{w}$  in order to minimise  $D$ .

Our adaptive weighting procedure is an iterative one, whereby at each iteration  $\tau$  we first update the previous weight vector  $\mathbf{w}^{(\tau-1)}$ , and then re-estimate  $\Pi^*(\mathbf{w}^{(\tau)})$  by fitting the GL model  $R$  times, each with a random permutation of the response in order to create  $R$  null data sets<sup>2</sup>.  $\Pi_l^*(\mathbf{w}^{(\tau)})$  is then the frequency at which pathway  $\mathcal{G}_l$  is selected across the  $R$  null data sets at iteration  $\tau$ . The algorithm is initialised at iteration  $\tau = 0$  by using an initial weight vector  $\mathbf{w}^{(0)}$ , for instance the standard size weighting (13). This procedure is then repeated until  $D$  reaches some suitably small value.

From (14), a reduction in  $D$  can be obtained by reducing the difference  $d_l = \Pi_l^*(\mathbf{w}) - \Pi_l$ , for all  $l$ . As each  $|d_l|$  approaches zero, the ratio,  $\Pi_l^*(\mathbf{w})/\Pi_l$ , approaches one, so that the contribution of pathway  $\mathcal{G}_l$  to  $D$  is decreased. With this in mind, at each iteration, we adjust pathway weights according to the following formula,

$$(15) \quad w_l^{(\tau)} = w_l^{(\tau-1)} [1 - \text{sign}(d_l)(\alpha - 1)L^2 d_l^2] \quad 0 < \alpha < 1$$

<sup>2</sup>Alternatively, in a simulation study where the null distribution of the response is known (as in section 3), the  $R$  models can be fitted after sampling a response from that null distribution.

where the parameter  $\alpha$  controls the maximum amount by which each  $w_l$  can be reduced in a single iteration, in the case that pathway  $\mathcal{G}_l$  is selected with zero frequency. The weighting update equation has the following desirable properties. When  $0 \leq \Pi_l^* < \Pi_l$ , i.e.  $-\frac{1}{L} \leq d_l < 0$ ,  $w_l$  is decreased, up to a maximum factor  $\alpha$  when  $\Pi_l^* = 0$ , increasing the chance that group  $l$  is selected. When  $\Pi_l^* > \Pi_l$ , i.e.  $d_l > 0$ ,  $w_l$  is increased, decreasing the chance that group  $l$  is selected. Finally, when  $\Pi_l^* = \Pi_l$ , i.e.  $d_l = 0$ ,  $w_l$  is unchanged. The square in the weight adjustment factor ensures that large values of  $|d_l|$  result in relatively large adjustments to  $w_l$ .

The estimation of  $\Pi^*$  when  $M > 1$ , that is where more than one pathway is selected by the model, is computationally infeasible even for a small value of  $M$ , since we would need to estimate the empirical *joint* probability distribution that  $M$  pathways are jointly selected. However, we expect that many of the factors biasing pathway selection when  $M = 1$  will similarly affect this joint probability distribution. Under this assumption, we estimate the optimal weight vector  $\mathbf{w}$  only in the  $M = 1$  case. Extensive simulation studies (see section 4) indicate that this data-driven adaptive weighting scheme is able to substantially increase power and specificity compared with the standard weighting (13), even when  $M > 1$ , indicating that this assumption holds in practice. Finally, we note that despite the need for multiple MC simulations over multiple iterations, our proposed bias-adjusted weighting strategy is fast, since it relies on fitting the GL model with  $\lambda$  tuned to select a single pathway only, ensuring that the active set (see section 2.2.3) is very small, and model estimation time for each of the  $R$  model fits is minimal.

## 2.4 Pathway ranking

Penalized regression typically proceeds by determining an optimal value for  $\lambda$ , corresponding to a subset of variables that best predicts the response, and this is generally done by cross validating the prediction error. In genetic association mapping, results are often instead presented in the form of lists of pathways or SNPs, ranked in order of importance. We seek such a strategy for the ranking of pathways within the regression model, such that pathways in  $\mathcal{C}$ , will achieve a high ranking, whereas those in  $\mathcal{C}'$  will be ranked low. This approach has the added advantage of allowing us to make direct comparisons with alternative pathway methods that use p-values as a ranking criterion.

One simple ranking criterion in penalised regression is to use the order in which each variable enters the model along the regularization path - i.e. as  $\lambda$  is decreased from its maximal value, where no variables are selected. We instead adopt a bootstrap sampling approach, in which we fit the regression model over multiple subsamples of the data, drawn with replacement, at a single, *fixed* value for  $\lambda$ . Pathways are ranked in order of importance according to their selection frequency across subsamples. Our motivation here is to exploit knowledge of finite sample variability obtained by subsampling, to achieve better estimates of pathway importance. In this respect our strategy resembles the pointwise stability selection method proposed by Meinshausen and Bühlmann (2010) in the context of variable selection.

As with stability selection, for our ranking strategy to be effective, the value of  $\lambda$  must be small enough to ensure that all pathways in  $\mathcal{C}$  are selected by the model with a high probability at each subsample. Computation time increases rapidly with  $M$ , the number of selected pathways, so that with the number,  $|\mathcal{C}|$ , of causal pathways unknown, the choice of  $M$  is driven by the number of causal pathways we seek to identify within computational constraints. We use  $B = 100$  subsamples, each of size  $N/2$ , and at each subsample we perform a line search over  $\lambda$ , to ensure that  $M \geq M_{min}$  pathways are selected. This procedure is described in appendix 5. Once  $\lambda$  is tuned, for each subsample,  $b$ , we obtain estimates  $\beta_j^{(b)} (b = 1, \dots, B)$  for each SNP coefficient ( $j = 1, \dots, P^*$ ). For pathway  $\mathcal{G}_l$ , we define  $\pi_l^{(b)} = 1$  when  $\|\beta_l^{(b)}\|_2 \neq 0$  and  $\pi_l^{(b)} = 0$  otherwise, where  $\beta_l^{(b)}$  is the pathway parameter vector estimated for subsample  $b$ . We rank pathways in order of their selection frequency across subsamples,  $\bar{\pi}_{l_1} \geq \dots \geq \bar{\pi}_{l_L}$ . We note that since typically  $M \ll L$ , some  $\bar{\pi}_l$  may be zero. Such pathways are classified as unranked.

## 2.5 Ranking performance measures

In order to evaluate the success of any PGAS method, some measure of ranking performance is required. In this section we describe 3 separate ranking performance measures that we use to evaluate the performance of our method in a simulation study described in section 3. One complicating factor is the issue of overlapping pathways, making the effective number of causal pathways,  $|\mathcal{C}|$ , dependent on the degree to which SNPs in  $\mathcal{S}$  overlap multiple pathways. In addition, with any method based on variable selection, the possibility that causal pathways are unranked, i.e. they are not selected by the model, must be taken into account.

Consider the situation where the set  $\mathcal{S}$  of causal SNPs, with cardinality  $S > 1$ , is known. We may choose to define  $\mathcal{C}$  in its most restricted sense as the set of pathways that contain *all* members of  $\mathcal{S}$ , or alternatively  $\mathcal{C}$  might include all pathways containing one or more SNPs belonging to  $\mathcal{S}$ . In either case  $|\mathcal{C}|$  will depend on the degree to which SNPs in  $\mathcal{S}$  overlap multiple pathways. This in turn depends on the particular distribution of causal SNPs with respect to overlapping genes. The need to accommodate this variability in  $|\mathcal{C}|$  in part motivates our formulation of the ranking measures described below.

We propose three separate ranking measures that capture different aspects of ranking performance, and focus on the top 100 ranked pathways only. We do this firstly because in any method attention is inevitably focused on the highest ranking pathways (or alternatively those with the highest statistical significance in a hypothesis testing framework). Also, since in a simulation study we compare the performance of our variable selection method which identifies a limited number of pathways against an alternative method that scores all pathways, some suitable cutoff in rank order must be chosen.

We denote the set of *ranked* causal pathways by  $\mathcal{C}^* = \{k \in \mathcal{C} : \bar{\pi}_k > 0\}$ , cardinality  $|\mathcal{C}^*|$ , and their respective rankings by  $r_{k_1}, r_{k_2}, \dots, r_{|\mathcal{C}^*|}$ , ranked in order of their respective selection frequencies,  $\bar{\pi}_{k_1} < \bar{\pi}_{k_2} < \dots < \bar{\pi}_{|\mathcal{C}^*|}$ . We further denote by  $\mathcal{C}_{100}^* = \{k \in \mathcal{C}^* : r_k \leq 100\}$ , cardinality  $|\mathcal{C}_{100}^*|$ , the set of ranked causal pathways falling in the top 100 ranks, with corresponding rankings  $r_{k_1}, r_{k_2}, \dots, r_{|\mathcal{C}_{100}^*|}$ . Our three proposed ranking measures are as follows:

1. *Highest causal pathway rank*,  $r_{k_1}$ , that is the single highest rank achieved by any pathway in  $\mathcal{C}_{100}^*$ . This lies in the range  $1 \leq r_{k_1} \leq 100$ , and is only defined for  $|\mathcal{C}_{100}^*| \geq 1$ .
2. *Ranking power*,  $p_{100}$ , defined as

$$(16) \quad p_{100} = \frac{|\mathcal{C}_{100}^*|}{|\mathcal{C}|}$$

with  $0 \leq p_{100} \leq 1$ .  $p_{100} = 0$  when no causal pathways are ranked in the top 100 ( $\mathcal{C}_{100}^* = \emptyset$ ), and  $p_{100} = 1$  when all causal pathways are ranked in the top 100 ( $\mathcal{C}_{100}^* = \mathcal{C}$ ).

3. *Power-adjusted, normalised, weighted ranking score*,  $R$ . This takes account of the actual rankings,  $r_{k_1}, \dots, r_{|\mathcal{C}_{100}^*|}$ , as well as the ranking power,  $p_{100}$ . We begin by defining a normalised, weighted ranking score,

$$(17) \quad R^* = \frac{\sum_{k \in \mathcal{C}_{100}^*} r_k^{\frac{1}{2}}}{\sum_{k=1}^{|\mathcal{C}_{100}^*|} k^{\frac{1}{2}}}$$

Here the square root increases the weight given to highly-ranked causal pathways. The denominator is a normalising factor which represents the minimum possible weighted ranking score, with  $r_{k_1} = 1, r_{k_2} = 2 \dots, r_{|\mathcal{C}_{100}^*|} = |\mathcal{C}_{100}^*|$ , ensuring that  $R^*$  attains its minimum value of 1 when the pathways in  $\mathcal{C}_{100}^*$  are optimally ranked. Higher values of  $R^*$  indicate suboptimal ranking.  $R^*$  takes no account of the possibility that  $\mathcal{C}_{100}^* \neq \mathcal{C}$ , i.e. not all causal pathways are ranked. To do this we form the adjusted measure

$$(18) \quad R = \begin{cases} R^*/p_{100} & \text{if } p_{100} > 0 \\ \gamma & \text{if } p_{100} = 0 \end{cases}$$

$R$  thus attains a minimum value of 1 when all causal pathways are optimally ranked, and the value  $\gamma$  when no causal pathways are ranked.

### 3 Simulation Study

We assess the power of our proposed method in a simulation study using real genotype and pathway data, with simulated, quantitative phenotypes generated under an additive genetic model from SNPs within a single, selected causal pathway. The presence of overlapping SNPs means that the actual number of causal pathways is typically greater than one. We additionally compare our method’s performance with an alternative, univariate-based method commonly used in gene set analysis. Computation times for both methods increase with  $P$ , and because of this, and the large number of scenarios and simulations tested, we restrict this analysis to SNPs on a single chromosome to keep execution times within practical limits.

#### 3.1 Genotype and pathways data

We use genotypes obtained from the Alzheimer’s Disease Neuroimaging Initiative, ADNI ([www.loni.ucla.edu/ADNI](http://www.loni.ucla.edu/ADNI)), derived from the Illumina Human 610-Quad BeadChip. Subjects comprise a mix of healthy controls, those diagnosed as having mild cognitive impairment, and those with AD. After removing variants with a call rate  $< 95\%$ , minor allele frequency (MAF)  $< 0.1$  and significant deviation from Hardy-Weinberg equilibrium ( $p < 5.7 \times 10^{-7}$ ), 448,294 SNPs remain. In this study we use genotype data from  $N = 743$  subjects, and consider only SNPs from chromosome 1 (33,850 SNPs).

Popular databases used for the mapping of genes to biological pathways include the Kyoto Encyclopedia of Genes and Genomes (KEGG, [www.genome.jp/kegg/pathway.html](http://www.genome.jp/kegg/pathway.html)) and BioCarta ([www.biocarta.com/genes/index.asp](http://www.biocarta.com/genes/index.asp)). For this study we use data on ‘canonical pathways’ from the Molecular Signals Database (MSigDB, [www.broadinstitute.org/gsea/msigdb/index.jsp](http://www.broadinstitute.org/gsea/msigdb/index.jsp)), which is a commonly-used, curated collection of pathways obtained from multiple sources. At the time of writing this comprised 880 pathways mapped to 6,804 genes. 2,382 human gene locations on chromosome 1, corresponding to assembly GRCh37.p3 are obtained using Ensembl’s biomart API ([www.biomart.org](http://www.biomart.org)). ADNI-genotyped SNPs on chromosome 1 are then mapped to annotated genes within 10kb (20,399 SNPs mapped to 2,096 genes), and these remaining genes and SNPs are then mapped to pathways using MSigDB (8,102 SNPs mapped to 778 pathways). Thus we see that the majority of chromosome 1 SNPs fail to map to any pathway, but that the majority of annotated pathways map to at least 1 SNP on this chromosome. Finally, small ( $< 10$  SNPs) and identical pathways are removed. After all pre-processing we are left with a total of  $P = 8,078$  SNPs mapped to 551 pathways (max: 1,059; min: 10; mean:  $120 \pm 142$  SNPs per pathway). All SNP to pathway mapping and filtering was performed using bespoke code written in Python. The mapping and filtering process is illustrated in Fig. 2.

More than 80% of SNPs are observed to overlap more than 1 pathway, with around 20% overlapping 10 or more pathways and 2% overlapping 60 or more (see Fig. 3). After variable expansion to account for overlapping pathways (section 2.1), we have  $P^* = 66,085$  SNPs.

#### 3.2 Simulation framework

We begin by adjusting the pathway weight vector,  $\mathbf{w}$ , using the bias-adjusted adaptive weighting procedure described in section 2.3. We do this over 10 iterations with  $R = 40,000$  MC simulations, each with response  $y$  sampled from a standard normal distribution,  $\mathcal{N}(0, 1)$  for simplicity, since many quantitative traits are expected to follow a normal distribution.

For the simulation of a SNP-dependent response, we begin by drawing  $S$  SNPs from a single, randomly selected causal pathway,  $\mathcal{G}_\phi$ , according to some specified distribution (see below), and then form the set  $\mathcal{C}$ , of causal pathways that contain all the members of  $\mathcal{S}$ . We thus chose to

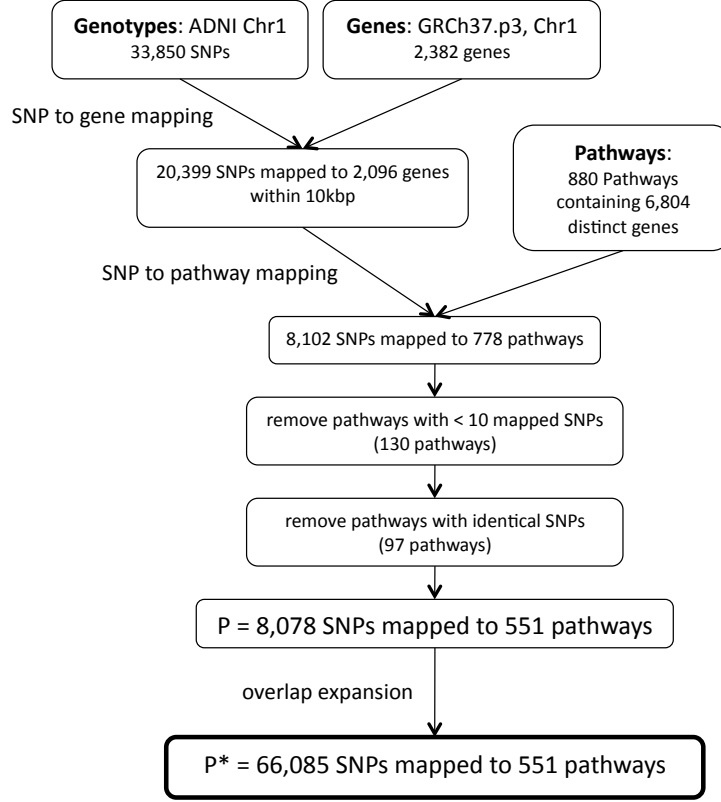


Figure 2: SNP to pathway mapping.

define  $\mathcal{C}$  in its most restricted sense, rather than for example including pathways that contain one or more SNPs in  $\mathcal{S}$ . Note that the number,  $|\mathcal{C}|$  of causal pathways will vary according to the particular distribution of overlaps within  $\mathcal{S}$ .

For each simulation, a univariate quantitative phenotype  $y$  is simulated using an additive model,

$$y_i = \sum_{k \in \mathcal{S}} \zeta_k x_{ik} + \epsilon$$

where  $\zeta_k$  is the allelic effect per minor allele due to causal SNP  $k$ . Setting  $w_k = \zeta_k x_k$ , we define the *effect size* of SNP  $k$  as  $\delta_k = E(w_k)/E(y)$  for  $k \in \mathcal{S}$ , and set  $\epsilon \sim \mathcal{N}(1, \sigma_\epsilon^2)$  so that  $\delta_k = 0$  when  $\zeta = 0$ . We also record the average SNP effect size as a proportion of total phenotypic variance,  $ES_k = \text{Var}(w_k)/\text{Var}(y)$ , and the mean proportionate change in response per minor allele,  $E(\zeta_k)$ . For our simulations we control  $\delta_k$ , and set  $\zeta_k$  accordingly, so that effect size is independent of SNP MAF, whereas  $\zeta_k$  and  $ES_k$  are MAF-dependent.

The power and specificity of any PGAS method is likely to depend on a range of factors including the number of causal pathways to be identified, the number and distribution of causal SNPs, and the size of their phenotypic effect (Wang et al., 2010; Fridley and Biernacka, 2011). We therefore assess the performance of our method across 6 different scenarios in which we vary each of these factors. Furthermore, we test each scenario over 500 MC simulations to account for variation in causal SNP MAFs, gene size and number within pathways, and LD patterns within and between causal pathways.

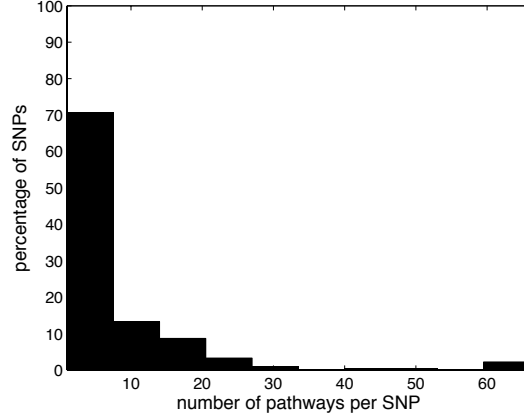


Figure 3: Frequency distribution of ADNI SNPs by number of pathways they map to. SNPs are mapped to genes within 10kbp. The data set consists of 8,078 SNPs and 551 pathways.

scenario	$S$	$\delta_k$	distribution	description
(a)	10	0.005	random from $\mathcal{G}_\phi$	$S$ large; $\delta_k$ large; random distribn
(b)	3	0.005	random from $\mathcal{G}_\phi$	$S$ small; $\delta_k$ large; random distribn
(c)	3	0.005	random from single gene in $\mathcal{G}_\phi$	$S$ small; $\delta_k$ large; single gene
(d)	10	0.001	random from $\mathcal{G}_\phi$	$S$ large; $\delta_k$ small; random distribn
(e)	3	0.001	random from $\mathcal{G}_\phi$	$S$ small; $\delta_k$ small; random distribn
(f)	3	0.001	random from single gene in $\mathcal{G}_\phi$	$S$ small; $\delta_k$ small; single gene

Table 1: Scenarios tested in simulation study. For scenarios (c) and (f), in the rare event that a gene has less than 3 SNPs, all SNPs within the gene are selected.

The list of scenarios tested is presented in Table 1. First, we consider scenarios where the number of causal SNPs is small ( $S = 3$ ) or large ( $S = 10$ ). Secondly, we consider two different SNP effect sizes. We choose values for  $\sigma_\epsilon^2$  and  $\delta_k$  to mimic effect sizes obtained in recent association studies, focussing particularly on the smallest reported effect sizes. Park et al. (2010) review GWAS for a number of quantitative traits (height, Crohn’s disease and breast, prostate and colorectal cancers) and report values for  $ES_k$  ranging from 0.02 to 0.0004. Cho et al. (2009) report values for  $\zeta_k$  for 8 quantitative traits in a large GWAS ranging from 1.6 to 0.006. A recent neuroimaging genetic study measuring genetic effects on a variety of traits related to brain structure reports significant values for  $\zeta_k$  of around 0.07 (Joyner et al., 2009). We set  $\sigma_\epsilon = 0.2$ , and test  $\delta_k = 0.005$  and 0.001, which gives values for  $ES_k = 0.001$  and 0.00004 and  $E(\zeta_k) = 0.01$  and 0.002 respectively. Finally, we also vary the particular distribution of SNPs with respect to their location within causal pathways. We expect the distribution of causal SNPs with respect to genes and associated LD blocks to affect performance, both in our regression model, and in the case where pathway scores are derived in a two-step process that begins with the calculation of gene association scores (Wang et al., 2007). The distributions of  $|\mathcal{C}|$ , the number of causal pathways for each scenario, are shown in Fig. 4.

## 4 Results

We begin with an investigation of the effect of our proposed speed ups to the GL estimation algorithm. We first note that GL estimation times will depend on the sample size ( $N$ ) and the number of SNPs ( $P$ ), which will in turn affect the number of mapped pathways ( $L$ ) and  $P^*$ . Estimation times will further depend on the number of groups selected ( $M$ ), and the amount of signal present, since this affects convergence times. For illustrative purposes, in Table 2 we show



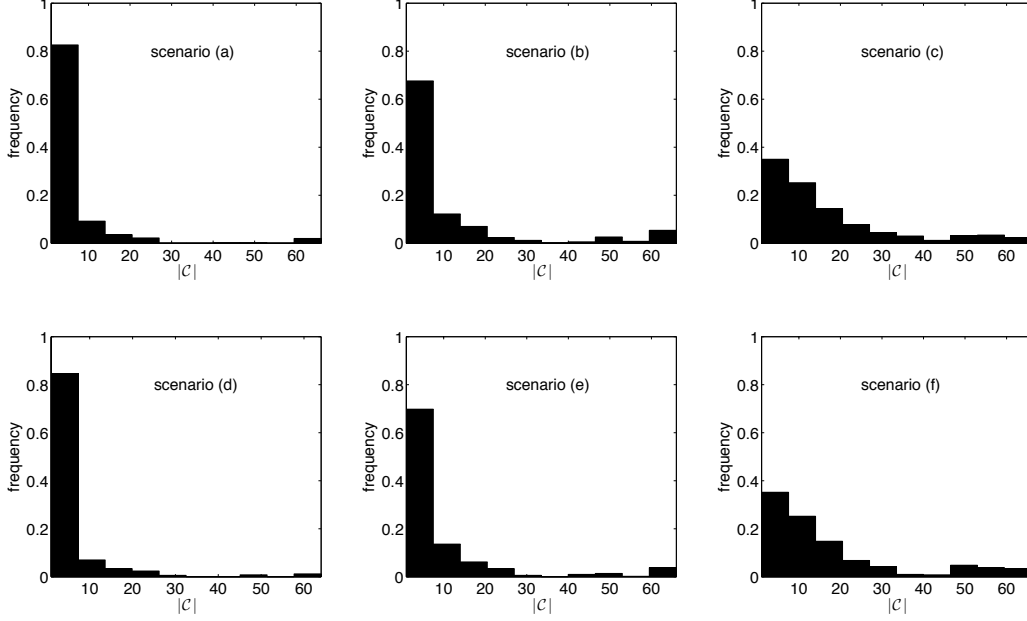


Figure 4: Distributions of  $|\mathcal{C}|$  across 500 MC simulations for the 6 scenarios described in Table 1. Where SNPs are distributed within a single gene (scenarios (c) and (f)), the number of causal pathways tends to be larger, since a single gene can map to multiple pathways. Where SNPs are distributed randomly across  $\mathcal{G}_\phi$  (scenarios (a), (b), (d), and (e)), this number tends to be smaller, particularly where the number of causal SNPs is large (scenarios (a) and (d)).

gains in execution time compared with ‘standard’ block coordinate descent, using our proposed speed ups for a single model fit with a null response, and for  $M = 10$ . Estimation times are seen to be substantially reduced across a range of values for  $N$  and  $P$ , dramatically so for larger datasets.

We next turn to the application of P-GLAW to real genotype and pathway data described in section 2.3. We apply this procedure over 10 iterations, each with  $R = 40,000$  MC simulations with a response  $y \sim \mathcal{N}(0, 1)$ . Fig. 5 (c) shows how the weight adjustment factor  $\mathbf{w}^{(\tau)}/\mathbf{w}^{(\tau-1)}$ , (see (15)), varies with  $d_l$  across all pathways at a single iteration. Fig. 5 (a) and (b) shows the observed, empirical distribution,  $\Pi^*$ , using the standard size weighting (13), and the adapted weights (15) after 10 iterations, respectively. The corresponding KL divergence measure,  $D$ , is observed to reduce steadily over the 10 iterations (Fig. 5 (d)), illustrating how the proposed weight adjustment procedure reduces pathway selection bias.

For the remainder of this section, we assess the performance of our proposed P-GLAW method using simulated phenotypes under the simulation framework described in the previous section, and using the bias-adjusted pathway weights described above. We first compare performance using the bias-adjusted weights with that obtained using the standard size weighting (13). We find the adjusted weighting scheme offers a considerable improvement in ranking performance for all ranking measures, and illustrate this in Fig. 6 for a single scenario (scenario (a)) using the ranking performance measures described in section 2.4. Fig. 6 (a) shows the first ranking measure ( $r_{k_1}$ ) as a ROC curve, in which we show the proportion of simulations with  $r_{k_1} \leq z$ , for ranks  $z = 1, 2, \dots, 100$ . We plot  $z$  on the horizontal axis as a false positive rate (FPR), so that  $\text{FPR} = (z - 1)/L$ . At a FPR of 0.05, we see that the adapted weighting scheme shows a more than 2 fold increase in power (from 0.29 to 0.62) over the standard pathway size weighting (13), indicating 62% of MC simulations have  $r_{k_1} \leq 28$ , compared with 29% for the standard size weighting. The distribution of  $p_{100}$  across 500 MC simulations is illustrated as a boxplot in Fig. 6 (b). Here we see that the adapted weighting scheme offers a clear and substantial improvement in

sample size	$P^* = 4k, L = 126$		$P^* = 66k, L = 551$		$P^* = 647k, L = 879$	
	BCD	BCD+	BCD	BCD+	BCD	BCD+
371 ( $N/2$ )	7.93	0.17	421	1.35	5490	16
743 ( $N$ )	16.9	0.27	511	2.5	6430	30.0

Table 2: GL estimation times (seconds) with  $M = 10$ . Table shows the time taken for the full estimation with a null  $\mathcal{N} \sim (0, 1)$  response, and with varying number of SNPs ( $P^*$ ), and sample size ( $N$ ). ‘BCD’ - estimation using block coordinate descent only. ‘BCD+’ - estimation using BCD with active set, Taylor approximation of the group penalty and efficient computation of block residuals. Genotype and pathway data as described in section 3.1.  $P^* = 4k$  : 5,000 SNPs from chromosome 1 mapped to 126 pathways.  $P^* = 66k$  : all 33,850 genotyped SNPs from chromosome 1 mapped to 551 pathways.  $P^* = 647k$  : 448,294 genome-wide SNPs mapped to 879 pathways. All computations performed using multi-threading on a single machine with 8 3.2 GHz processors and 64GB RAM.

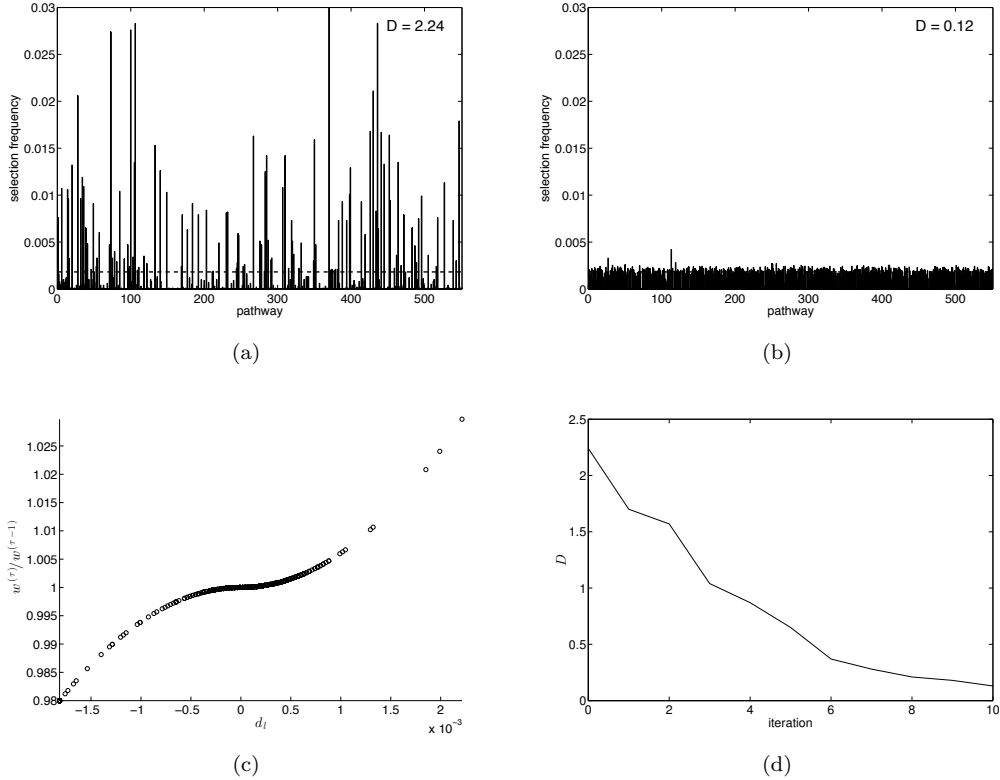


Figure 5: Application of bias-adjusted weighting procedure to the data used in the simulation study.  $R = 40,000$ , with a different null response,  $y \sim \mathcal{N}(0, 1)$ , at each MC simulation.  $\alpha = 0.98$ . (a) Empirical pathway selection frequency distribution,  $\Pi^*$ , with standard, pathway size weighting,  $w_l = \sqrt{S_l}$ .  $D = 2.24$ . Dotted horizontal line shows the expected distribution,  $\Pi_l = 1/L \simeq 0.002$ . (b)  $\Pi^*$  with bias-adjusted weights after 10 iterations.  $D = 0.12$ . (c) Variation of weighting adjustment factor  $w^{(\tau)}/w^{(\tau-1)}$  with  $d_l$  at a single iteration, with  $\alpha = 0.98$ . Each point represents the adjustment to a single  $w_l, l = 1, \dots, L$ . (d) Decrease in K-L divergence,  $D$ , over 10 iterations.

GL’s capacity to rank a high proportion of causal pathways in the top 100 ( $p = 2.03 \times 10^{-50}$  that the two population  $p_{100}$  CDFs are equal using a two-sample Kolmogorov-Smirnov (KS) test). GL with the standard weighting scheme performs particularly poorly with 55% of simulations failing to rank any causal pathway in any simulation, compared with 18% for the adapted weighting scheme. Finally, Fig. 6 (c) shows the distribution of the  $R$  ranking measure across 500 simulations under the two weighting schemes. Once again we see that the adaptive weighting scheme demonstrates improved ranking performance over the standard size weighting scheme, with the distribution of  $R$  scores skewed towards lower values for the former, indicating that causal pathways tend to be ranked higher.

We next assess P-GLAW ranking performance with the adapted weighting scheme across the full range of scenarios, and compare these with pathway rankings obtained using the method proposed by Wang et al. (2007), commonly referred to as ‘GenGen’ (GG). GG is a widely-used, GSEA-type PGAS method that measures pathway enrichment using genes scores derived from univariate SNP statistics. Studies using GG include searches for implicated pathways in Crohn’s disease (Wang et al., 2009b), autism spectrum disorders (Wang et al., 2009a), breast cancer (Menashe et al., 2010) and Alzheimer’s disease (Lambert et al., 2010). GG begins by scoring each SNP according to its association with the phenotype. SNPs are then mapped to genes within a specified distance, and each gene is scored according to its most significant mapped SNP. The enrichment of highly-ranked genes in a given pathway is then compared with those in all other pathways, to obtain a pathway enrichment score. For GenGen we use identical source data (genotypes, phenotypes, SNP to gene, and gene to pathway mappings), and rank pathways by normalised enrichment score, determined from 1,000 permutations (the GG default settings). MC simulations for P-GLAW and GG are performed in parallel across 50 (P-GLAW) and 500 (GG) processors respectively, on a high-performance computing cluster. As described above for alternative weighting schemes, results for the comparison study are presented in the form of  $r_{k_1}$  ROC curves (Fig. 7),  $p_{100}$  boxplots (Fig. 8) and  $R$  bar graphs (Fig. 9). Selected ranking measures are presented in numerical form in Tables 3 and 4.

scen.	ROC power, fpr = 0.05			median $p_{100}$			proprn. $p_{100} = 0$			KS 2 sample test $p_{100}$ cdfs the same
	P-GLAW	GG	ratio	P-GLAW	GG	ratio	P-GLAW	GG	ratio	
(a)	0.62	0.35	1.76	0.60	0.60	1.00	0.18	0.26	0.70	p = 0.0082
(b)	0.61	0.33	1.84	0.33	0.11	3.00	0.21	0.45	0.46	p = $9.6 \times 10^{-25}$
(c)	0.81	0.54	1.49	0.35	0.20	1.73	0.06	0.23	0.25	p = $2.5 \times 10^{-25}$
(d)	0.44	0.18	2.37	0.33	0.00	$\infty$	0.30	0.62	0.48	p = $7.7 \times 10^{-27}$
(e)	0.59	0.27	2.18	0.33	0.01	37.33	0.23	0.50	0.46	p = $9.2 \times 10^{-28}$
(f)	0.79	0.45	1.74	0.31	0.14	2.31	0.06	0.31	0.20	p = $3 \times 10^{-38}$

Table 3: Selected ranking performance measures for P-GLAW and GG for the 6 scenarios described in Table 1. *ROC power, fpr = 0.05*: proportion of 500 MC simulations with  $r_{k_1} \leq 28$  corresponding to a fpr of 0.05. *median  $p_{100}$* : median of  $p_{100}$  distribution across 500 MC simulations. *Proportion with  $p_{100} = 0$* : proportion of 500 MC simulations with no causal pathway in the top 100 ranks. *KS 2 sample test*: two-sample Kolmogorov-Smirnov test of the hypothesis that the P-GLAW and GG  $p_{100}$  population cdfs are the same.

Beginning with the ROC curves illustrating the  $r_{k_1}$  ranking measure (Fig. 7 and first 3 columns of Table 3), GG consistently demonstrates increased power and specificity across all of the top 100 ranks illustrated. In addition, the relative gain in power for P-GLAW is greater at the smallest effect size for each equivalent scenario, (a) vs. (d), (b) vs. (e), and (c) vs. (f). At the smaller effect size, where causal SNPs are distributed randomly within causal pathways, power increases where the number of causal SNPs is fewer ((d) vs. (e)). Finally, maximum power is achieved for both methods where causal SNPs are located within a single gene ((c) and (f)).

Turning to the distributions of the  $p_{100}$  ranking measure (Fig. 8, and columns 4 to 9 in Table 3), P-GLAW again outperforms GG across all scenarios. For example, the null hypothesis that the two population cdfs are equal is rejected at the  $\alpha = 0.05$  level (Table 3, final column), as is the null hypothesis that the two sample medians are the same (Fig. 8), except for scenario (a) where

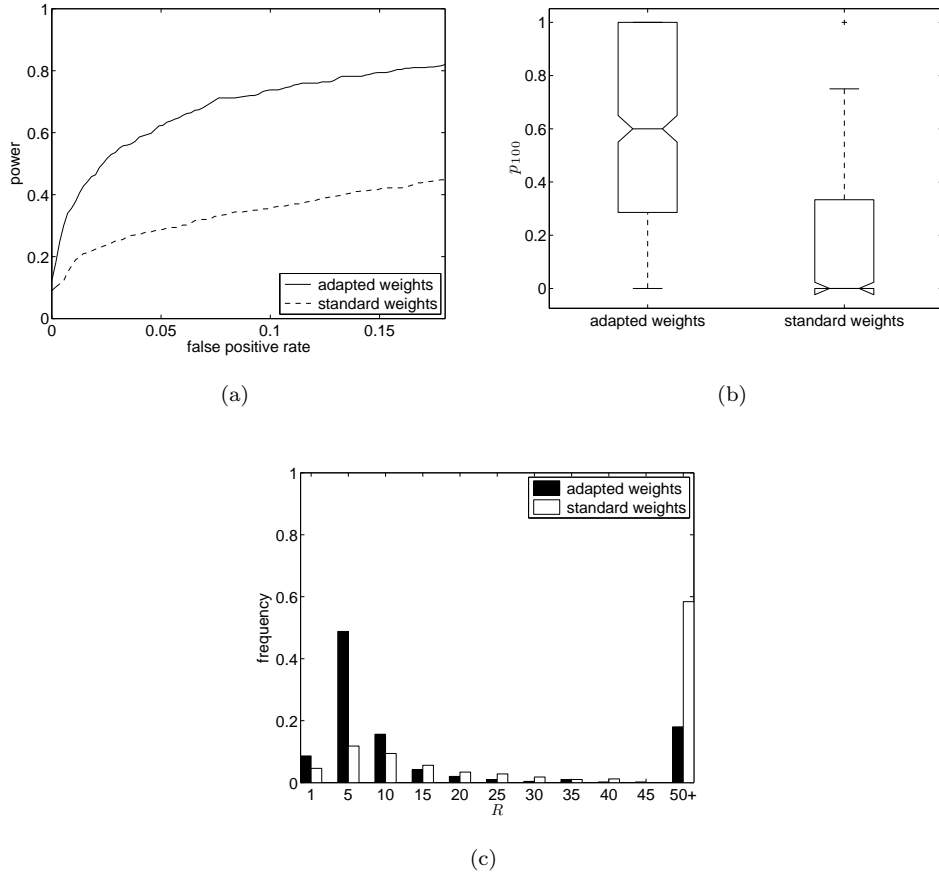


Figure 6: Comparison of ranking performance: adaptive weighting scheme (section 2.3) vs. standard pathway size weighting (13).  $S = 10$ ;  $\delta_k = 0.005$ ; SNPs randomly distributed across  $\mathcal{G}_\phi$ . (a) ROC curves illustrating power to identify at least one causal pathway in the top 100. Power is average across 500 simulations. (b) Distribution of ranking power,  $p_{100}$ , across 500 simulations. This is the proportion  $|\mathcal{C}_{100}^*|/|\mathcal{C}|$  of causal pathways in  $\mathcal{C}$  that are ranked in the top 100 pathways. Notches indicate 95% confidence intervals for the true median. (c) Distribution of the power-adjusted, normalised, weighted ranking score,  $R$ , across 500 simulations. The final '50+' column includes simulations for which no causal pathway was ranked in the top 100, i.e.  $\mathcal{C}_{100}^* = \emptyset$ ;  $R = 100$ .

median  $p_{100}$  is not significantly different for the two methods. Excluding scenario (a) where both methods perform relatively well, P-GLAW median  $p_{100}$  is consistent across each scenario, and is maintained from the larger to the smaller effect size. This is in marked contrast to GG, where this measure shows a large decrease at the smaller effect size, although the decrease is less marked when causal SNPs are located within a single gene. A similar pattern persists for both P-GLAW and GG if we consider the proportion of simulations with  $p_{100} = 0$ , i.e. where no causal pathways are found in the top 100 ranks, except for P-GLAW in the case where causal SNPs are located in a single gene, where this measure is particularly low.

The final series of plots (Fig. 9), illustrate the distributions of  $R$  across all scenarios. These distributions once again follow the trends in ranking performance highlighted above, but they offer a more nuanced view, in the sense that while this measure takes power into account, it is also sensitive to the actual causal pathway rankings. Here we see that P-GLAW tends to rank causal pathways higher than GG, since all P-GLAW distributions are skewed towards lower  $R$  values, indicating that causal pathways tend to be ranked higher. This is borne out if we focus on the proportion of simulations with  $R < 10$  (Table 4, first 3 columns), which also illustrates how proportionate gains in ranking performance for P-GLAW over GG are largest for the smallest effect size ((a)-(c) vs. (d)-(f)). This table also gives results for the proportion of simulations demonstrating near optimal ranking of causal pathways ( $R < 3$ ), although the very small frequencies suggest that little can be inferred from these.

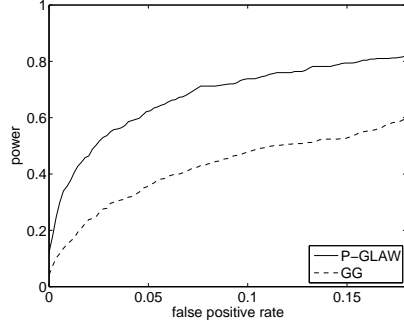
scenario	$R < 10$			$R < 3$		
	P-GLAW	GG	ratio	P-GLAW	GG	ratio
(a)	0.68	0.46	1.47	0.13	0.09	1.38
(b)	0.50	0.24	2.11	0.03	0.03	0.93
(c)	0.55	0.33	1.68	0.01	0.07	0.18
(d)	0.44	0.20	2.22	0.03	0.02	2.00
(e)	0.46	0.20	2.33	0.02	0.03	0.69
(f)	0.45	0.23	1.96	0.01	0.04	0.30

Table 4: Proportion of 500 simulations with  $R < 10$  and  $R < 3$  for the 6 scenarios described in Table 1.

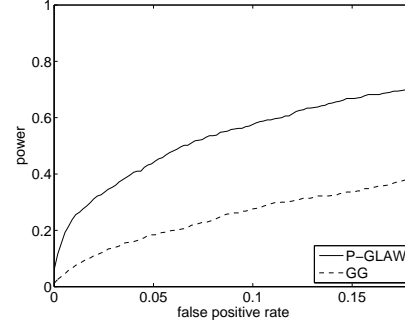
## 5 Discussion

We have developed a penalised regression-based strategy (P-GLAW) that exploits functional structure within genotypes to identify biological pathways associated with a continuous trait. We use the group lasso, with all mapped SNPs and pathways in a single regression model, and use a novel combination of methods including a bias-adjusted group weighting scheme and bootstrap sampling, together with a number of speed ups designed to make the analysis of large scale datasets computationally feasible. An important feature of our method is the need to accommodate the presence of overlapping pathways. On the assumption that causal SNPs are enriched within a biological pathway, we find in a simulation study that our proposed method shows relative gains in both power and specificity across a range of scenarios, compared with an alternative pathways method (GG), based on univariate SNP statistics, that we use as a benchmark. We believe this is the first such study evaluating GL performance using real SNP and pathway data across a range of realistic scenarios.

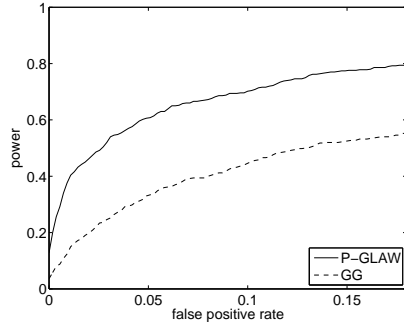
One key motivation for a pathways-based approach is the desire to harness the joint effects of those SNPs or genes with relatively small effect size, that typically fail to achieve genome-wide significance in GWAS (Baranzini et al., 2009). We hypothesise that the advantages inherent in a multivariate approach to modelling SNP effects will increase power to detect these, and in our simulation study we therefore focus on scenarios with causal SNPs that exhibit effect sizes at or below the limits of those found in recent GWAS. To evaluate the performance of each method considered here, we devise three separate ranking metrics, each of which measures a different aspect of ranking performance.



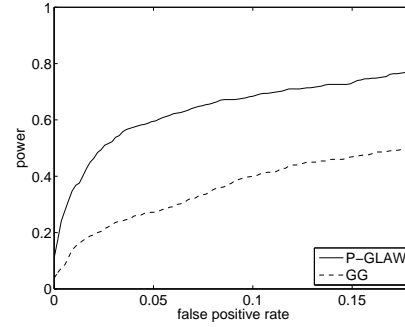
(a)  $S = 10; \delta_k = 0.005$ ; random distbn



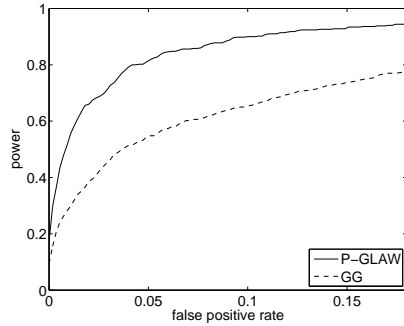
(d)  $S = 10; \delta_k = 0.001$ ; random distbn



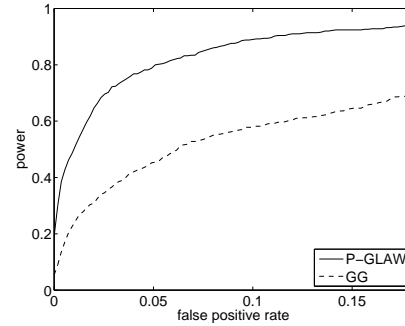
(b)  $S = 3; \delta_k = 0.005$ ; random distbn



(e)  $S = 3; \delta_k = 0.001$ ; random distbn

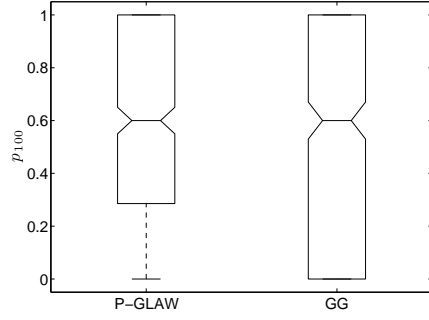


(c)  $S = 3; \delta_k = 0.005$ ; single gene distbn

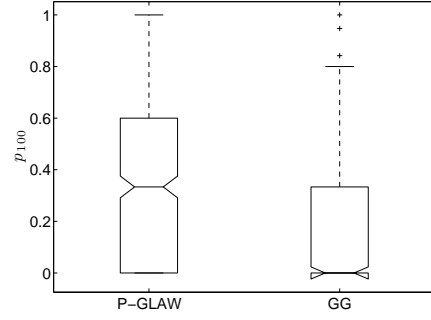


(f)  $S = 3; \delta_k = 0.001$ ; single gene distbn

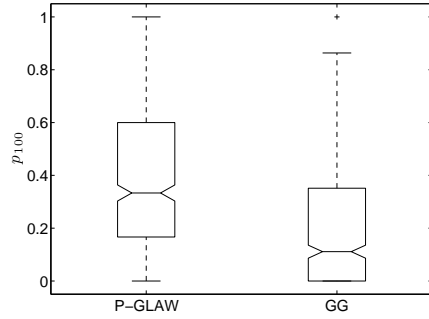
Figure 7: ROC curves illustrating proportion of simulations with  $r_{k_1} \leq z$ , for ranks  $z = 1, 2, \dots, 100$ . Power is average across 500 simulations. False positive rate  $= (z - 1)/L$ . Scenarios corresponding to the higher SNP effect size ( $\delta_k = 0.005$ ) are presented in the left-hand column, with the equivalent scenarios at the lower effect size ( $\delta_k = 0.001$ ) on the right.



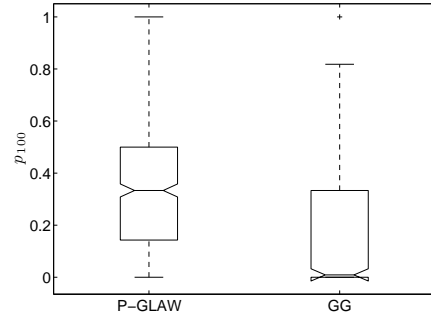
(a)  $S = 10; \delta_k = 0.005$ ; random distbn



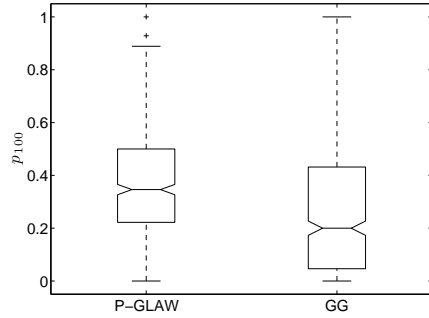
(d)  $S = 10; \delta_k = 0.001$ ; random distbn



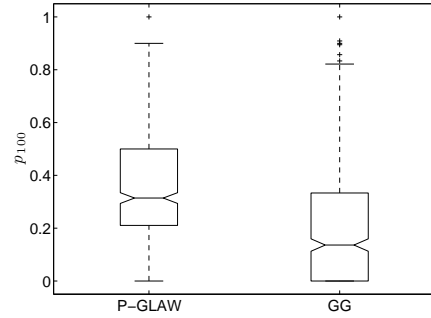
(b)  $S = 3; \delta_k = 0.005$ ; random distbn



(e)  $S = 3; \delta_k = 0.001$ ; random distbn

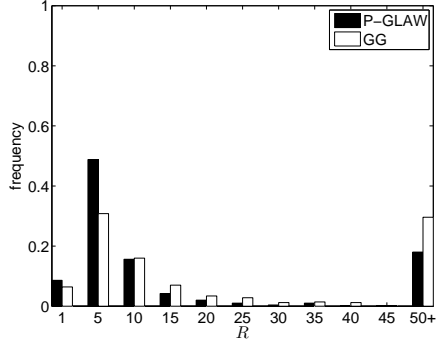


(c)  $S = 3; \delta_k = 0.005$ ; single gene distbn

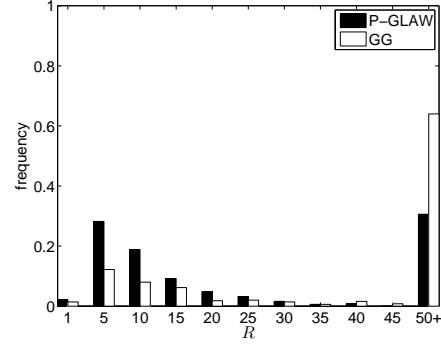


(f)  $S = 3; \delta_k = 0.001$ ; single gene distbn

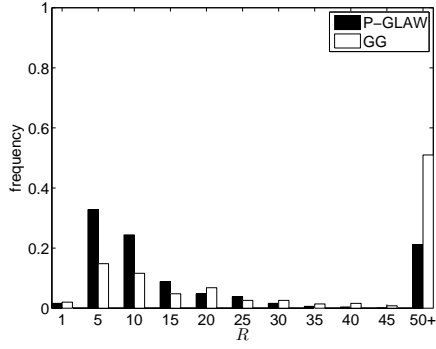
Figure 8: Box plots of distribution of ranking power,  $p_{100}$ , across 500 simulations. This is the proportion  $|\mathcal{C}|_{100}^*/|\mathcal{C}|$  of causal pathways in  $\mathcal{C}$  that are ranked in the top 100 pathways. Notches indicate 95% confidence intervals for the true median.



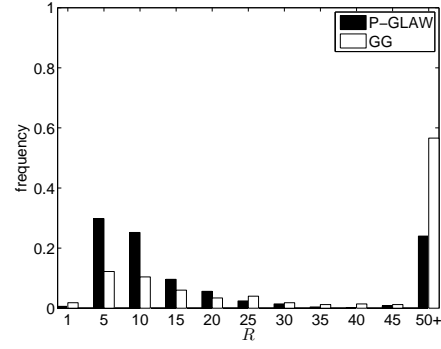
(a)  $S = 10; \delta_k = 0.005$ ; random distbn



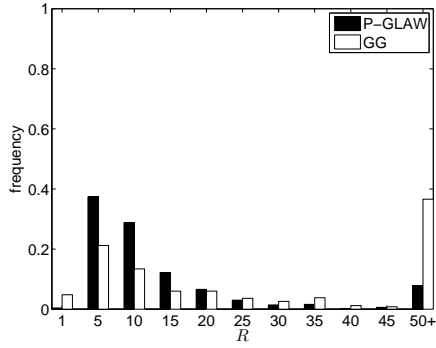
(d)  $S = 10; \delta_k = 0.001$ ; random distbn



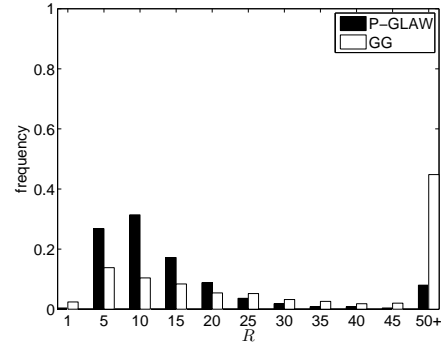
(b)  $S = 3; \delta_k = 0.005$ ; random distbn



(e)  $S = 3; \delta_k = 0.001$ ; random distbn



(c)  $S = 3; \delta_k = 0.005$ ; single gene distbn



(f)  $S = 3; \delta_k = 0.001$ ; single gene distbn

Figure 9: Distribution of the power-adjusted, normalised, weighted ranking score,  $R$ , across 500 simulations. The final '50+' column includes simulations for which no causal pathway was ranked in the top 100, i.e.  $\mathcal{C}_{100}^* = \emptyset$ ;  $R = 100$ .



One factor affecting power is the ‘genetic architecture’ of the disease in question, that is the number and distribution of SNP effects across causal pathways (Wang et al., 2010). For example, causal SNPs may be distributed across many genes in a pathway, or restricted to a single gene. Since PGAS methods vary in the way that they combine the effects of individual SNPs, the specific genetic architecture is expected to impact power for different methods in different ways (Wang et al., 2009b; Holmans et al., 2009). GG uses genes scores corresponding to the most significant SNP associated with a gene to establish pathway significance. This has the advantage of reducing redundant information arising from SNPs in LD with a causal SNP within a single gene, but may lead to reduced power where causal variants reside in distinct LD blocks within a gene (Wang et al., 2007). An important, related factor that we find has received little attention is the issue of overlapping pathways, and the consequent effect on PGAS performance. The precise distribution of causal SNPs with respect to genes that overlap multiple pathways will affect the number of pathways that are considered to be ‘causal’, and we expect this to affect ranking performance for different methods in different ways. To explore these issues, we investigate a variety of different genetic architectures, in which we vary both the number and distribution of causal SNPs with respect to pathways and genes.

In general, we find that P-GLAW performs well across the range of causal SNP distributions and effect sizes considered. Additionally, our method is able to consistently outperform the benchmark (GG). GG performance at the smaller effect size is particularly weak, so that P-GLAW shows the largest gains in relative performance here.

An insight into some of those factors affecting ranking performance is afforded by considering some of the ranking measures in more detail. Starting with the highest ranking causal pathway measure ( $r_{k_1}$ ), as expected we find that this decreases for each scenario at the smaller effect size. However, at the smaller effect size this measure is observed to increase for both methods as the number of causal SNPs is decreased, markedly so when the reduced number of causal SNPs are concentrated in a single gene. Since the effect size for each causal SNP is held constant, this seems counterintuitive, since the pathway ‘signal’ is reduced when there are fewer causal SNPs. In addition, for the reasons described above, for GG this signal may be further reduced where causal SNPs reside within a single gene. The explanation is likely to be that the effective number of causal pathways tends to increase as the number of causal SNPs is reduced, increasing the probability that a single causal pathway is ranked high. The number of causal pathways is even larger when causal SNPs are concentrated in a single gene (see Fig. 4). Where the pathway signal is highest (scenario (a)), both methods tend to rank a high proportion of causal pathways in the top 100 (high  $p_{100}$ ), although the proportion of MC simulations in which GG fails to rank any causal pathways (that is the proportion of simulations with  $p_{100} = 0$ ) is relatively high. On this measure of ranking power, GG performs relatively poorly across all other scenarios, particularly at the smaller effect size. Interestingly, P-GLAW is relatively insensitive to variation in the number and distribution of SNPs within causal pathways, as might be expected from the smoothing properties of the GL  $\ell_2$  penalty, which ensures that all SNPs within a selected pathway are retained in the model (Zhou et al., 2010).

The need to account for factors such as variation in LD, gene and pathway size is a feature common to all PGAS methods. A range of approaches, often used in combination, have been proposed to correct for these biasing factors, including the use of gene scores that summarise SNP statistics (Holmans et al., 2009), and permutation of phenotypes (Wang et al., 2009b). Dimensionality reduction techniques have also been advocated for the control of redundant information (Chen et al., 2010; Zhu and Li, 2011; Ballard et al., 2010). For P-GLAW, we propose a method that adjusts the distribution of pathway weights according to the observed bias in pathway selection frequencies across multiple MC simulations under the null. We find in a simulation study that our proposed bias correction method does substantially increase power and specificity, indicating that pathway selection bias is decreased. One potential disadvantage of our approach is that it takes no account of variation in biasing factors within a pathway. It would be interesting to compare the relative merits of our approach against alternative bias-reduction methods, for example the use of within-pathway dimensionality reduction. However, we consider the retention of all SNPs in the regression model to be a potentially attractive feature of our approach, as it affords the

possibility of the simultaneous identification of causal SNPs driving pathway selection, and we are currently pursuing this as an extension to the present model.

In situations where predictors, or groups of predictors are correlated, both the lasso and group lasso can demonstrate problems with consistency, that is they are unable to constantly identify the true set of causal predictors or groups (Zhao and Yu, 2006; Bach, 2008; Chatterjee and Lahiri, 2011). Despite this, we have demonstrated that in a finite sample, our bootstrap sampling approach performs well, and this has been borne out elsewhere (Meinshausen and Bühlmann, 2010). We are however pursuing alternative methods for the ranking of pathways, using different ranking strategies.

We pay considerable attention to the need to develop fast algorithms for solving the GL, a problem that is particularly acute when using regression models with GWAS data. Using a combination of techniques, we establish a GL estimation algorithm that can quickly solve the GL using whole genome data. However, the very large number of simulations and scenarios considered in our simulation study, and the relatively slow performance of the benchmark method mean that we restrict the analysis to mapped SNPs from a single chromosome.<sup>3</sup>

We note that phenotypes in our simulation study are generated under an additive linear model. The assumption of additive linear SNP effects is built into both the P-GLAW and GG models, in the former through the SNP allele codings in the genotype design matrix, and in the latter through the particular model used to generate the univariate SNP scores, although for both methods alternative models can easily be accommodated.

In our simulation study we account for variation in the size and distribution of causal SNP minor allele frequencies through the use of MC simulations, but we expect that such variation is likely to impact model performance, and this is something that warrants further exploration.

As with all PGAS methods, we note that results are dependent on the choice of pathways database, and will inevitably reflect biases due for example to the increased number of annotations for genes implicated in particular disease etiologies (Elbers et al., 2009; Cantor et al., 2010). Results are also subject to bias resulting from SNP to gene mapping strategies. For example, SNP to gene mapping distances will affect the number of unmapped SNPs falling within gene ‘deserts’ (Eleftherohorinou et al., 2009), SNPs will map to relatively large numbers of genes in gene rich areas of the genome, and the mapping of a SNP to its closest gene may obscure a true functional relationships with a more distant gene (Wang et al., 2009b).

Finally, we note that our method can be easily adapted to accommodate other ways of grouping SNP data, for exampling using protein interaction networks (Wu et al., 2010), or GO and other ontologies (Jensen and Bork, 2010).

## Appendix Line search over $\lambda$

We wish to tune  $\lambda$  so as to select a minimum  $M$  pathways at each subsample. To do this we perform a line search over  $\lambda$ . This procedure is described in box 3.

**Box 3** Line search procedure for tuning  $\lambda$  to select  $M \geq M_{min}$  pathways

1. Set  $\lambda_{max} = \min_{\lambda} : \|\mathbf{X}_l^T \mathbf{y}\|_2 \leq \lambda w_l$  (from (12)) and  $\alpha = 0.8^\dagger$
2. Let  $\lambda = \alpha \lambda_{max}$
3. Form the active set,  $\mathcal{A} = \{m \in \mathcal{G} : \|\mathbf{X}_m^T \mathbf{y}\|_2 \leq \lambda w_m\}$
4. Let  $M = |\mathcal{A}|$ . If  $M < M_{min}$  skip to step 6.<sup>‡</sup>
5. Solve the GL estimation at  $\lambda$  using the active set  $\mathcal{A}$ , as described in box 2 (starting at box 2, step 2.)

---

<sup>3</sup>Python code for mapping SNPs to pathways, and for analysing SNP data using PGLAW is available on request.

Let the solution be  $\hat{\beta}$ , with final active set  $\mathcal{A}$   
 $\mathcal{S}(\lambda) = \{l \in \mathcal{G} : \|\hat{\beta}_l\| > 0\}$  (the set of selected pathways)  
 $M = |\mathcal{S}(\lambda)|$  (the number of selected pathways)

6. if  $M \geq M_{min}$   
 $\hat{\beta}$  is the full solution  
STOP  
else  
 $\lambda_{max} = \lambda$  (need to decrease  $\lambda$ )  
end
7. Go to step 2.

<sup>†</sup> The value of  $\alpha$  is chosen for computational convenience. A value close to 1 ensures that  $\lambda$  values stay close to 1, so that as few pathways are selected by the model as possible, thus speeding up the estimation. However, a value too close to 1 means that the decrease in  $\lambda$  at each iteration is small, meaning that many iterations may have to be performed before  $M$  reaches the desired range.

<sup>‡</sup> This step is introduced for computational efficiency, since if  $|\mathcal{A}| < M_{min}$  there is no prospect of selecting enough groups

## References

- Bach, F. (2008): “Consistency of the group lasso and multiple kernel learning,” *J. Mach. Learn. Res.*, 9, 1179–1225.
- Ballard, D. H., J. Cho, and H. Zhao (2010): “Comparisons of multi-marker association methods to detect association between a candidate region and disease.” *Genetic epidemiology*, 34, 201–12.
- Baranzini, S. E., N. W. Galwey, J. Wang, P. Khankhanian, R. Lindberg, D. Pelletier, W. Wu, B. M. J. Uitdehaag, L. Kappos, C. H. Polman, P. M. Matthews, S. L. Hauser, R. A. Gibson, J. R. Oksenberg, and M. R. Barnes (2009): “Pathway and network-based analysis of genome-wide association studies in multiple sclerosis.” *Human molecular genetics*, 18, 2078–90.
- Bigos, K. L. and D. R. Weinberger (2010): “Imaging genetics-days of future past.” *NeuroImage*, 53, 804–809.
- Breheny, P. and J. Huang (2009): “Penalized methods for bi-level variable selection,” *Statistics and Its Interface*, 2, 369–380.
- Cantor, R. M., K. Lange, and J. S. Sinsheimer (2010): “REVIEW Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application,” *American Journal of Human Genetics*, 86, 6–22.
- Chatterjee, A. and S. Lahiri (2011): “Bootstrapping Lasso Estimators,” *Journal of the American Statistical Association*, 106, 608–625.
- Chen, L. S., C. M. Hutter, J. D. Potter, Y. Liu, R. L. Prentice, U. Peters, and L. Hsu (2010): “Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data,” *American Journal of Human Genetics*, 86, 860–871.
- Cho et al. (2009): “A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits.” *Nature genetics*, 41, 527–34.
- Elbers, C. C., K. R. van Eijk, L. Franke, F. Mulder, Y. T. van der Schouw, C. Wijmenga, and N. C. Onland-Moret (2009): “Using genome-wide pathway analysis to unravel the etiology of complex diseases.” *Genetic epidemiology*, 33, 419–31.
- Eleftherohorinou, H., C. J. Hoggart, V. J. Wright, M. Levin, and L. J. M. Coin (2011): “Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways,” *Human molecular genetics*, 1–13.
- Eleftherohorinou, H., V. Wright, C. Hoggart, A.-L. Hartikainen, M.-R. Jarvelin, D. Balding, L. Coin, and M. Levin (2009): “Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases.” *PloS one*, 4, e8068.

- Fan, J. and R. Li (2001): "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.
- Fridley, B. L. and J. M. Biernacka (2011): "Gene set analysis of SNP data: benefits, challenges, and future directions." *European journal of human genetics : EJHG*, 19, 837–843.
- Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani (2007): "Pathwise coordinate optimization," *Annals of Applied Statistics*, 1, 302–332.
- Friedman, J., T. Hastie, and R. Tibshirani (2010): "A note on the group lasso and a sparse group lasso," Available at <http://www-stat.stanford.edu/~tibs/ftp/sparse-grlasso.pdf>, 1–8.
- Goldstein, D. B. (2009): "Common genetic variation and human traits." *The New England journal of medicine*, 360, 1696–8.
- Hoggart, C. J., J. C. Whittaker, M. De Iorio, and D. J. Balding (2008): "Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies." *PLoS genetics*, 4, e1000130.
- Holmans, P., E. K. Green, J. S. Pahwa, M. a. R. Ferreira, S. M. Purcell, P. Sklar, M. J. Owen, M. C. O'Donovan, and N. Craddock (2009): "Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder." *American journal of human genetics*, 85, 13–24.
- Jacob, L., G. Obozinski, and J.-p. Vert (2009): "Group Lasso with Overlap and Graph Lasso," in *Proceedings of the 26th International Conference on Machine Learning*.
- Jensen, L. J. and P. Bork (2010): "Ontologies in quantitative biology: a basis for comparison, integration, and discovery." *PLoS biology*, 8, e1000374.
- Joyner, A. H., C. R. J., C. S. Bloss, T. E. Bakken, L. M. Rimol, I. Melle, I. Agartz, S. Djurovic, E. J. Topol, N. J. Schork, O. A. Andreassen, and A. M. Dale (2009): "A common MECP2 haplotype associates with reduced cortical surface area in humans in two independent populations," *Proceedings of the National Academy of Sciences*, 106, 15483–15488.
- Lambert, J.-C., B. Grenier-Boley, V. Chouraki, S. Heath, D. Zelenika, N. Fievet, D. Hannequin, F. Pasquier, O. Hanon, A. Brice, J. Epelbaum, C. Berr, J.-F. Dartigues, C. Tzourio, D. Campion, M. Lathrop, and P. Amouyel (2010): "Implication of the immune system in Alzheimer's disease: evidence from genome-wide pathway analysis." *Journal of Alzheimer's disease : JAD*, 20, 1107–18.
- Lango Allen et al. (2010): "Hundreds of variants clustered in genomic loci and biological pathways affect human height," *Nature*, 467, 832–838.
- Lesnick, T. G., S. Papapetropoulos, D. C. Mash, J. Ffrench-Mullen, L. Shehadeh, M. de Andrade, J. R. Henley, W. a. Rocca, J. E. Ahlskog, and D. M. Maraganore (2007): "A genomic pathway approach to a complex disease: axon guidance and Parkinson disease." *PLoS genetics*, 3, e98.
- Luo, L., G. Peng, Y. Zhu, H. Dong, C. I. Amos, and M. Xiong (2010): "Genome-wide gene and pathway analysis." *European journal of human genetics : EJHG*, 18, 1045–1053.
- Ma, S. and M. R. Kosorok (2010): "Detection of gene pathways with predictive power for breast cancer prognosis." *BMC bioinformatics*, 11, 1.
- Manolio et al. (2009): "Finding the missing heritability of complex diseases," *Nature*, 461, 747–753.
- Meinshausen, N. and P. Bühlmann (2010): "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 417–473.
- Menashe, I., D. Maeder, M. Garcia-Closas, J. D. Figueroa, S. Bhattacharjee, M. Rotunno, P. Kraft, D. J. Hunter, S. J. Chanock, P. S. Rosenberg, and N. Chatterjee (2010): "Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade." *Cancer research*, 70, 4453–9.
- Mootha, V. K., C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop (2003): "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes." *Nature genetics*, 34, 267–73.
- Park, J.-H., S. Wacholder, M. H. Gail, U. Peters, K. B. Jacobs, S. J. Chanock, and N. Chatterjee (2010): "Estimation of effect size distribution from genome-wide association studies and implications for future discoveries." *Nature genetics*, 42.
- Plomin, R., C. M. a. Haworth, and O. S. P. Davis (2009): "Common disorders are quantitative

- traits." *Nature reviews. Genetics*, 10, 872–878.
- Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira, D. Bender, J. Maller, P. Sklar, P. De Bakker, and M. Daly (2007): "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," *The American Journal of Human Genetics*, 81, 559–575.
- Roth, V. and B. Fischer (2008): "The Group-Lasso for Generalized Linear Models : Uniqueness of Solutions and Efficient Algorithms," in *Proceedings of the 25th International Conference on Machine Learning*.
- Tibshirani, R. (1996): "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58, 267–288.
- Tibshirani, R., J. Bien, J. Friedman, T. Hastie, S. Noah, and J. Taylor (2010): "Strong Rules for Discarding Predictors in Lasso-type Problems," Available at <http://arxiv.org/abs/1011.2234>.
- Tseng, P. and S. Yun (2009): "A coordinate gradient descent method for nonsmooth separable minimization," *Mathematical Programming*, 117, 387–423.
- Vounou, M., T. E. Nichols, G. Montana, and A. D. N. I. (2010): "Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach." *Neuroimage*, 53, 1147–1159.
- Wang et al. (2009a): "Common genetic variants on 5p14.1 associate with autism spectrum disorders." *Nature*, 459, 528–33.
- Wang, K., M. Li, and M. Bucan (2007): "Pathway-based approaches for analysis of genomewide association studies." *American journal of human genetics*, 81, 1278–83.
- Wang, K., M. Li, and H. Hakonarson (2010): "Analysing biological pathways in genome-wide association studies," *Nature Reviews Genetics*, 11, 843–854.
- Wang, K., H. Zhang, S. Kugathasan, V. Annese, J. P. Bradfield, R. K. Russell, P. M. A. Sleiman, M. Imielinski, J. Glessner, C. Hou, D. C. Wilson, T. Walters, C. Kim, E. C. Frackelton, P. Lionetti, A. Barabino, J. V. Limbergen, S. Guthery, L. Denson, D. Piccoli, M. Li, M. Dubinsky, M. Silverberg, A. Griffiths, S. F. A. Grant, J. Satsangi, R. Baldassano, and H. Hakonarson (2009b): "Diverse Genome-wide Association Studies Associate the IL12 / IL23 Pathway with Crohn Disease," *American Journal of Human Genetics*, 84, 399–405.
- Wu, G., X. Feng, and L. Stein (2010): "A human functional protein interaction network and its application to cancer data analysis." *Genome biology*, 11, R53.
- Wu, T. T., Y. F. Chen, T. Hastie, E. Sobel, and K. Lange (2009): "Genome-wide association analysis by lasso penalized logistic regression." *Bioinformatics (Oxford, England)*, 25, 714–21.
- Yuan, M. and Y. Lin (2006): "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49–67.
- Zhao, J., S. Gupta, M. Seielstad, J. Liu, and A. Thalamuthu (2011): "Pathway-based analysis using reduced gene subsets in genome-wide association studies." *BMC bioinformatics*, 12, 17.
- Zhao, P. and B. Yu (2006): "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.
- Zhou, H., M. E. Sehl, J. S. Sinsheimer, and K. Lange (2010): "Association Screening of Common and Rare Genetic Variants by Penalized Regression." *Bioinformatics (Oxford, England)*, 26, 2375–2382.
- Zhu, H. and L. Li (2011): "Biological pathway selection through nonlinear dimension reduction." *Biostatistics (Oxford, England)*, 429–444.
- Zou, H. and R. Li (2008): "One-step Sparse Estimates in Nonconcave Penalized Likelihood Models." *Annals of statistics*, 36, 1509–1533.